DEALING WITH MISSING DATA ON ALCOHOL CONSUMPTION USING DIET DIARIES IN A BIRTH COHORT STUDY

Margaret Ely

Thesis submitted for the Degree of Doctor of Philosophy
of the University of London

Department of Epidemiology and Public Health
University College London

2004

UMI Number: U602551

All rights reserved

INFORMATION TO ALL USERS
The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript
and there are missing pages, these will be noted. Also, if material had to be removed,
a note will indicate the deletion.

UMI U602551
Published by ProQuest LLC 2014. Copyright in the Dissertation held by the Author.
Microform Edition © ProQuest LLC.
All rights reserved. This work is protected against
unauthorized copying under Title 17, United States Code.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

ABSTRACT

Recent alcohol research has focussed on the importance of patterns of drinking rather than on total consumption over a period. This requires collection of detailed data, as in a daily diary, with a resulting tendency for a substantial proportion of missing data. In the past, dealing with missing data in epidemiology was based mainly on naïve methods.

The aim of this dissertation is to critically examine ways of dealing with missing data on alcohol consumption collected in diet diaries by the 1946 birth cohort study, and to develop a method which takes account of both the technical statistical problems which arise with such data and the characteristics of the data which are of substantive importance in alcohol research. Recent developments in standard statistical software packages (SPSS, S-Plus), and special-purpose packages for missing data analysis (such as SOLAS™), have given epidemiologists access to more sophisticated approaches such as propensity score, linear regression, EM algorithm and methods of multiple imputation. These methods are evaluated using a simulation-based approach, which demonstrates that ignoring missing data, or handling them incorrectly, can lead to inefficient and biased results. A technical problem arises because the distribution of alcohol consumption is semicontinuous. The results show some standard methods are not suitable for variables of this kind, some use inappropriate algorithms, whilst others are not appropriate for epidemiological research because they do not preserve relationships between variables. Single or deterministic imputation methods fail to take account of uncertainty about the missing values.

The thesis shows how, using Schafer's procedures for multiple imputation, the information in alcohol diary data can be fully exploited and efficient inferences made. The multiply imputed datasets can be used for any subsequent analysis. Examples used in this thesis are the prevalence of excessive alcohol consumption, the role of alcohol consumption in the relationship between birthweight and blood pressure in mid-life and the dependence of blood pressure on alcohol consumption.

Any method of dealing with missing data should evaluate the sensitivity of inferences to its assumptions. In this thesis the sensitivity of inferences to the MAR assumption and to the model for imputation is evaluated.

Acknowledgements

I would like to thank Professor Mike Wadsworth for his encouragement, Dr Nick Longford for his nice attention to the drafts of this thesis, Dr Rebecca Hardy for her practical advice and guidance and Dr Alison Paul for her help with access to the raw diet diary data and the reproduction of Appendix 2.

I would like to extend my thanks to Ted Harding for not allowing me to give up and for his painstaking proof reading and typesetting of this manuscript; and to Alice and Jess for their forbearance in the absence of their mother.

Last, but not least, I would like to thank all the members of the NSHD, without whose contribution this study would not have been possible.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Introduction

Alcohol consumption is one of the major public health issues of our time (Royal College of Psychiatrists, 1986) but its measurement poses a problem for epidemiological studies. Epidemiological studies require information about individual levels of alcohol consumption in the general population. This information is derived from surveys in which a sample of subjects from the population is asked what they drink. In practice, the epidemiologist using survey data has at his disposal a sample of people in which not everyone who was designed to take part does so, or in which those who do take part do not provide all the requested information. In general, the precise reasons for this non-response are not known. It is commonly believed that people who drink excessively are less likely to be represented in general population surveys than those who drink moderately or not at all. It is thought that heavier drinkers are less likely to be contacted because they change address more frequently or are homeless, or are more likely to refuse an interview (Lemmens et al., 1988). In addition they may be less likely to complete items on their drinking habits, or, if they do, they may be more likely to under-report their drinking because of the associated stigma or forgetting (Smyth and Browne, 1991). As a result epidemiological studies of alcohol consumption in the general population may be biased.

Traditionally epidemiologists have paid cursory attention to the problem of missing data. The standard approach has been to analyse only cases with complete data. This is not only an inefficient use of data since information contained in partially complete records is discarded, but gives biased estimates if the non-respondents differ systematically from the respondents. Problems of bias are often dealt with only qualitatively in the discussion of results. Meanwhile, the statistical theory of missing data has advanced through the work of Little and Rubin (1987), who proposed multiple imputation as a general solution. The methodology of multiple imputation provides a way of enabling statistically valid inferences to be made using all available information, including that contained in incomplete data. Until recently the practical application of multiple imputation has been limited to those with statistical programming expertise. However the work of Rubin (1987) increased the awareness of survey analysts to the issue of missing data. Procedures specifically dealing with missing data became included in standard software packages in the 1990s. For example, SPSS released its 'Missing Value Analysis' procedures in 1996. A specialist software package for dealing with missing data, SOLAS, was first released by Statistical Solutions in 1997. The first release of SOLAS (1.0) included a method which used multiple imputation, Propensity Score, whilst v2.0, released in

1999, added model based methods for multiple imputation. Schafer, a student of Rubin, was developing his software for multiple imputation in 1996, but this was only in part available as a stand alone-package for Windows from 1999. Schafer's methods were released as part of S-Plus 6 in 2001.

This dissertation investigates methods of dealing with missing data, in the context of epidemiological research on alcohol consumption. It uses data on alcohol consumption derived from diet diaries in a birth cohort study, the MRC National Survey of Health and Development (NSHD). It examines the implications of using multiple imputation of missing data on alcohol consumption for inferences about the prevalence of excessive alcohol consumption (marginal distribution of an outcome variable with missing values), the role of alcohol consumption in the relationship between birthweight and blood pressure in mid-life (missing values in a potential confounder) and the dependence of blood pressure on alcohol consumption (missing values in the independent variable).

1.2 Difficulties of measuring alcohol consumption

1.2.1 The need to measure alcohol consumption

> *Alcohol in the clinical model may be merely a device; when an epidemiologist considers a large population, he observes that it is alcohol which causes alcoholism.*
>
> (Kessel and Walton, 1989).

In the earlier part of the twentieth century, alcohol research focused on alcoholism as a disease suffered by unfortunate individuals. Within the theoretical framework of this 'disease' or medical model, research focused on clinical studies of those receiving treatment (Royal College of Psychiatrists, 1986; Kessel and Walton, 1989). By definition they drank too much, and so what they drank was not seen as the focus of the problem for research. Research was focused on the factors associated with psychological dependency, the psychological 'weaknesses' of the individual, and negative precursors in childhood such as parental mistreatment or sexual abuse (Kessel and Walton, 1989; Edwards, 1994; Plant, 1997). Alcoholics were seen as disordered personalities. However, the epidemiologist interested in public health is concerned with the whole population (Edwards, 1994). Increasing alcohol consumption in the post-war era has been associated with increasing levels of alcohol related harm, or at least awareness of it (Royal College of Psychiatrists, 1986). Alcohol consumption, rather than the individual, was seen as the agent of the harm, since the availability of alcohol, rather than the number of disordered personalities, had increased (Kessel and Walton, 1989; Rose, 1992). The problems created by alcohol consumption were viewed as the responsibility of society rather than the individual. It was recognised that problems could be caused by drinking in individuals who would not be considered to be dependent on alcohol, and at levels of consumption below those associated with alcoholism (Royal College of Psychiatrists, 1986; Edwards, 1994). A *Lancet* editorial in 1977 suggested that *'the bulk of alcohol induced damage is in fact being experienced by non-dependent drinkers whose troubles do not resemble the medical stereotype of alcoholism.'*

(p. 1087). Certainly large scale population studies of liver cirrhosis in France suggested that the level of 'safe' drinking is much lower than previously thought (Tuyns et al., 1983). In this, the 'social model', drink is the agent and alcohol consumption by individuals in the general population is of interest.

Recent alcohol research has focussed on the importance of patterns of drinking rather than on total consumption over a period (Grant and Litvak, 1997), and this requires more detailed data (Grand and Single, 1997). Increasingly, there is an interest in how different patterns of consumption are related to social and health consequences. Rehm et al. (1996) noted that 'Patterns of drinking introduce the social element into alcohol epidemiology'. In a review of evidence about drinking patterns and their consequences, Rehm et al. conclude that social harm and casualties seem to be more closely linked than chronic health conditions to patterns of drinking. By drinking patterns they mean the way drinking is distributed over the week as opposed to the average quantity drunk. Average quantity is most closely linked theoretically with chronic diseases (Rehm et al., 1996), whilst social harm or accidents and injury are thought to be linked to binge drinking, that is, an excessive amount on an occasion. It is believed that the drinking pattern most relevant for health outcome is the distinction between binge drinking and sustained drinking (Arria and Gossop, 1997). There is also an historical link with occupational social class and drinking style in men, typified in a study in London in the 1960s (Edwards et al., 1972). This work seems to justify the stereotypes of the working class drinking man bingeing in the pub at the weekend, the upper class drinking man taking his whisky after lunch in the club, or wine with dinner at home in the evening. Do such stereotypes exist at the end of the 20th Century? The *BMJ* quotes Marmot on the effect of drinking: *'how it increases wife battering and falls from building sites'* (Dillner, 1995). Stephen Dorrell, the then Minister for Health, speaking on *Panorama* (BBC TV) stated that *'the majority of whisky is drunk by people who have no alcohol problem and who have no health problem arising from alcohol'.*

Patterns of drinking may refer to several aspects of drinking behaviour, including temporal variations in drinking, the number of heavy drinking occasions, the settings where drinking takes place, the activities associated with drinking, the types of beverage consumed, and the clusters of drinking norms and behaviours often referred to as drinking cultures (Single and Leino, 1997). Across Europe national drinking patterns are quite different. The Mediterranean style is characterised by regular daily drinking, often with meals, predominantly wine. In comparison the Nordic style of binge drinking is concentrated at weekends and predominantly of beer or spirits. Recent studies have indicated a convergence of drinking patterns in Europe: Mediterranean countries drinking more beer and Northern European countries drinking more wine; while 'wet' countries become drier and 'dry' ones wetter (Hupkens et al., 1993).

1.2.2 The problems of measuring alcohol consumption

The measurement of individual alcohol consumption poses a major problem for alcohol research. Average drinking levels of the population in Britain and Western Europe could be

estimated from the excise duty paid on alcoholic beverages. Such data has been available in Britain for the past 300 years (Spring and Buss, 1997). However, such estimates are thought to underestimate consumption because they do not include alcohol consumed but on which no duty is paid such as from private distilleries or smuggled imports, and, more recently, from the increase in duty free drinks brought into the country from the Common Market (Kessel and Walton, 1989). However, what does the average level for the population tell us about the drinking levels of individuals? It was argued, from international comparative studies, that increasing total consumption in a population has a direct positive relationship with adverse consequences, such as liver cirrhosis mortality (Pequignot, 1980). The relationship between national consumption and mortality from particular causes is straightforward to examine in the western world where records are kept of national alcohol sales and causes of death. It was inferred that there was a positive relationship between mean alcohol consumption in a population and the proportions of those drinking above a certain limit. This has been empirically demonstrated in many studies (see Edwards, 1994). However, the relationship between mean and proportion drinking excessively is not a necessary one, as it is possible that the mean of a distribution rises because light drinkers drink more, whilst heavy drinkers do not. Ledermann (1956) proposed that there was a precise (mathematical) relationship between the mean and the proportion above a limit, because alcohol consumption followed a precisely specifiable distribution. Although the logic of his conclusions can be refuted (Duffy, 1993), his work supported the 'social' model and provided the background for public health policies on 'sensible' consumption during the 1980s.

1.2.3 Survey instruments for measuring alcohol consumption

If we cannot deduce the alcohol consumption of individuals from national data, then we have to use general population surveys which ask people how much they drink. Estimates based on survey data can dramatically underestimate total consumption even compared with the estimates based on excise duty (Pernanen, 1974). Alcohol consumption is subject to measurement error. Ideally we would want to be able to estimate lifetime consumption. The best approximation might be obtained from prospective cohort studies with frequent measurement of alcohol consumption. Such a design is rarely found in alcohol studies, because of the expense of following a whole population for so long. Specialist alcohol studies are generally retrospective, and suffer from recall bias. In Britain there has only been an interest in the alcohol consumption of individuals in general population studies since the 1980s. The National Food Survey (Ministry of Agriculture Fisheries and Food), conducted from the 1940s until 1995, included alcohol consumption per household only and also excluded drinking outside the home. The ONS General Household Survey has included questions about drinking every two years from 1978 (Office for National Statistics, 1999). The Health and Lifestyle Survey, conducted in 1984/5 and 1991/2, included several measures of alcohol consumption (Cox et al., 1993). From 1991, the Health Survey for England carried out an annual survey on behalf of the Department of Health including data on respondents' estimated usual weekly alcohol consumption (Prescott-

Clarke et al., 1998). The MRC National Survey of Health and Development, a cohort study of people born in Britain in 1946, first asked about their alcohol consumption in the 1980s, when the members were in midlife, and the National Child Development Study (1958 British Birth Cohort) did so at ages 23 and 33.

Even if we restrict our focus to current or recent drinking, there is the problem of how to summarise this variable quantity. One possibility is to ask about the 'usual' amount drunk (*quantity* of drink) and the *frequency* of drinking, called a *Q-F index.* For example, prior to 1986, the General Household Survey (GHS) asks how often respondents drank different types of drink during the previous year, and how much of each they usually drank on one occasion. This is used to estimate drinking over a period, such as a week, by a process of averaging. However, people do not necessarily drink identical or similar quantities every week, and averaging results in error particularly for those whose dominant cycle of drinking does not coincide with the period reported (Uchtenhagen, 1990). The GHS stopped using the Q-F index in 1986 as it was thought to be misleading: in particular, it was believed to underestimate consumption by people in non-manual groups, who are more likely to drink different types of drink than those in manual groups (Wilson, 1980). Since 1986 the GHS measured alcohol consumption based on people's estimates of 'amounts usually drunk on any one day', which still means the subject themselves must 'average' their consumption in some way, unless they regularly drink the same amount daily. Such measures require a subjective assessment by each respondent, which may be influenced by their attitudes to drinking, their perception of the purpose of the survey, or the mode of administration of the survey instrument (self report or interviewer; postal or telephone) (Aquilino, 1992).

Another approach is to ask respondents about the actual amount drunk during a specific time period. This may be a summary total which is recalled — for example the subject may be asked to recall the total number of alcoholic drinks they have had in the past week. It is feasible to recall a total only over a short period unless the respondent is a very regular drinker, and a short period is less likely to be representative. For example, Duffy and Alanko (1992) found that a considerable number of individuals who were classified as light or heavy drinkers based on a summary of their previous week's drinking, are actually moderate drinkers if assessed over a longer period. With retrospective measures the longer the period asked about the better the representation (because of including those with longer drink cycles), but the greater the potential for recall bias.

It has been found that the more specific the measure, that is the closer a measure comes to assessing actual quantities consumed, the more reliable it is (Alanko, 1981). A more specific measure of alcohol consumption can be obtained by asking subjects to record the actual amount drunk on each drinking occasion, or to keep a *diary* of their drinking over a period of time, for example one week. A daily diary has the advantage of avoiding recall bias (Leigh, 2000). Self-reported alcohol consumption derived from surveys is generally found to underestimate

consumption compared with national data on the sale of alcohol derived from excise duty returns, by as much as 40 to 60 percent (Kessel and Walton, 1989). This phenomenon is known as undercoverage. The diary has been shown to give greater coverage of sales data than a weekly recalled amount or Q-F index, which suggests it has a higher validity (Lemmens et al., 1992).

In addition, a diary allows different aspects of drinking behaviour, or patterns of drinking, to be identified, not subjectively but by the analyst. In the same way that average national alcohol consumption cannot represent the diversity of people's drinking, the average or total alcohol consumption of an individual cannot adequately represent their drinking behaviour. Recent alcohol research has focussed on the importance of patterns of drinking rather than on total consumption over a period (Grant and Litvak, 1997), and this requires more detailed data (Grand and Single, 1997). The diary method of collecting data on alcohol consumption is able to capture different aspects of drinking pattern: it can yield not only total alcohol consumption, but also, for example, frequency of drinking, excessive daily drinking or frequency of heavy drinking.

In summary using a prospective diary to measure alcohol consumption has several advantages over other methods. It avoids recall bias or forgetting and minimises bias arising from under-reporting in subjective assessments or summaries of quantity drunk, whilst giving more detailed data that provide a richer source of data on patterns of consumption. However, it has the disadvantage of demanding a greater commitment from the subjects, who are more likely to drop out or to fail to complete the diary.

1.3 Dealing with missing data

1.3.1 Types of missing data

Inference is the process of generalising from a sample to a population: and assumes that the sample is representative of the population. The sample may not be a simple random selection from the population. For example, subjects may be randomly selected in different proportions, depending on their membership of particular subsets of the population (stratified sampling), or with probabilities proportional to some quantitative attribute (sampling proportional to size). Since the selection is made by a known probabilistic mechanism, which is incorporated in the sampling design, such inequalities can be adjusted for in the inference. For the individuals who happen not to be selected, their data are of course not intended to be observed, and hence are missing, but they are missing by design, according to a known mechanism. In addition, some data which are intended to be observed may be missing and, in this case, the reason for the data being missing is generally not known: this is called non-response and is essentially of two types. First, not everyone who was selected to take part in a survey may do so (perhaps because they are unwilling or because they are unavailable). This is called 'case non-response'. Second, those who do take part may not provide all the information intended to be collected. This is called 'item non-response'.

1.3.2 Problems posed by missing data

Where cases are missing by design, the probabilistic properties are controlled, and we can adjust for them in the statistical analysis. However, non-response (where values are missing not by design) poses a threat to representativeness of the available data. Unless it can be assumed that non-response occurs completely at random, the available data values may no longer be representative of the population values, and estimates based on the available data may be biased.

The simplest way to deal with item non-response is to ignore cases with incomplete data, using only cases with complete data, an approach called complete case analysis. There are two problems with is approach: it reduces the efficiency of estimates because the number of cases included is reduced, and the estimates are biased if those who complete diaries (responders) are systematically different from those who do not (non-responders). This approach is valid, yielding unbiased estimates, only under the assumption that the responders are a random sub-sample of the population being represented. In other words, it is valid only if non-response occurs completely at random within the sample, a condition known as missing completely at random (MCAR). The assumption is that non-responders are no different from responders.

It is, however, not possible to prove that data are missing completely at random. Supporting evidence for MCAR is often provided by comparing the distribution of observed baseline variables in responders and non-responders. If, it is argued, the responders are generally representative of the whole sample or population with respect to individual baseline variables then there is unlikely to be any non-response bias. For example, there may be no significant difference between the proportion of men and women in the responders and non-responders. There are three problems with this argument. First the non-significance does not guarantee no difference because there may be insufficient power to detect the difference. Secondly, variables which are univariately independent of a response variable may not be jointly independent. For example, suppose women and men are equally likely to respond to alcohol consumption questions, as are married and unmarried subjects. It is possible that married women are less likely to respond than unmarried women while married men are more likely to respond than unmarried men. Thirdly, subjects with incomplete data may be systematically different from those with complete records. Even though women and men may be equally likely to complete alcohol consumption questions, it is possible that women who do not respond may be heavier drinkers than those who do.

On the other hand, we have only to find one observed characteristic that is related to non-response to undermine the MCAR assumption. For example, respondents in higher social classes may be more likely to complete their diaries than those in lower social classes. If, in addition, those in lower social classes drink more excessively than those in higher social classes, then the overall distribution of alcohol consumption in responders will be biased. A qualitative approach, which has often been used in epidemiological research, is to infer the direction of the bias of an estimate based on assumptions that are thought to be plausible. In the above example,

we could argue that the sample proportion of people drinking excessively, based on the observed data, underestimates the true level, or is downwardly biased. The tacit assumption of this argument is that the relationship between alcohol consumption and social class is the same for non-responders as for responders. The usual expression for this assumption is that alcohol consumption is missing at random (MAR) conditional on social class. Assessing the impact of any bias due to non-response in this qualitative way is only practical in the simplest of analyses, and does not give us a quantitative assessment of bias. To properly evaluate the sensitivity of our inferences to the MAR assumptions, we need to explore the quantitative impact of simulations which embody these assumptions. One way of doing this is by imputation.

1.3.3 Dealing with item non-response by imputation

The MAR assumption makes it possible to exploit the information contained in incomplete records. We can do this explicitly by using the relationships in the observed data to model the missing values as functions of the background variables. We assume that the values are missing at random conditional on the background variables. This enables us to explore quantitatively the impact of the assumption on inferences. Item non-response can be dealt with by filling in a plausible value for each missing data item, a process called imputation. The common sense approach, which relies on the MAR assumption, is to look at the observed characteristics of those with missing data to provide some information about their missing response. If someone does not report their alcohol consumption we would look at other known characteristics which are indicative of alcohol consumption. For example, alcohol consumption is likely to be greater if the survey respondent is a man rather than a woman, or a smoker rather than a non-smoker. The characteristics that are related to the variable of interest provide background information. The relationship between the variable with missing data and other background variables can be modelled using the observed data. An advantage of imputation is that once the missing data values have been filled in the completed dataset can be analysed using standard complete-data methods.

1.3.4 Multiple imputation and its advantages

Multiple imputation creates a number (m) of imputations for each missing value. The first set of imputation values is used to form the first dataset, and so on, so that m completed datasets are obtained. Each of the m completed datasets is analysed using standard complete-data statistical methods, and the results are combined using a simple algorithm. (Rubin and Schenker, 1991; given here in Section 2.9). A method for imputation should:

1 Take into account the uncertainty about the imputed values

2 Be efficient—exploit as much of the information in partially complete records as possible

3 Preserve the observed relationships between the variables with missing values and the background variables.

These conditions will be shown to be fulfilled by maximum likelihood methods for multiple imputation both in theory (Sections 2.7–2.8) and in practice by testing different methods on simulated data (Chapter 6).

The problem remains that subjects with incomplete data may be systematically different from those with complete records, over and above what can be accounted for by models based on background data. In other words, the MAR assumption may not hold. Indeed, it seems plausible that people who drink heavily would be less likely to provide information about it, i.e. that the amount of alcohol consumed is likely to influence the chance of non-response. In such a situation, alcohol consumption is missing not at random (MNAR). The MAR assumption cannot be tested empirically (as can the MCAR assumption). It will be argued that more naïve methods of imputation fail to exploit all the information in partially complete records. For this reason, multiple imputation has a greater potential to avoid the problem of MNAR, as will be argued below. An example given earlier is the assumption that alcohol consumption is MAR conditional on social class. We assumed that the relationship between alcohol consumption and social class is the same for non-responders as for responders. Now this may not be a valid assumption. Suppose that, unknown to the imputer, heavy-drinking lower social class men are more likely to complete their diaries while heavy-drinking higher social class men are less likely to complete their diaries. Then, conditional on social class (and gender), alcohol consumption is not MAR. However, suppose the reason for this behaviour is that higher social class men feel guilty about heavy drinking, and feeling guilty is associated with non-response, while lower social class men do not feel guilty about heavy drinking. Suppose also that we know whether people feel guilty about their drinking. If the relationship between feeling guilty, social class and alcohol consumption in men is observed, then alcohol consumption is MAR conditional on guilt and social class (and gender). MAR fails to hold only if alcohol consumption depends on the missing values after conditioning on the available information. The more information we have about people's drinking, the more likely it is that we can predict the alcohol consumption of non-responders. The assumption of MAR becomes more plausible the richer the set of conditioning variables (Schafer, 1997).

1.3.5 The sensitivity of inferences to assumptions about missing data

The aim of the epidemiologist is to make valid inferences. The validity of inferences based on multiple imputation depends on the following assumptions:

1 The mechanism of missingness (MAR)

2 The model for the imputation

and also

3 The model for the analysis

Since the assumptions about the missing data (1 and 2 above) cannot be proven, it is important to acknowledge our uncertainty about them by performing sensitivity analysis. The original

philosophy of multiple imputation was that imputed datasets could be used in any subsequent analyses. Even when both the processes of imputation and of analysis can be implemented by the epidemiologist, the question remains whether the imputation can be used in any subsequent analysis, thereby avoiding the problem of dealing with the missing data again. It is feasible, but is it sensible to have a 'once and for all' imputation of alcohol consumption? There has been recent controversy over the validity of multiple imputation inferences when the imputer's and analyst's models differ (Fay, 1992; Meng, 1994; Rubin, 1996). The validity of the inferences derived using the rules for combining the multiple imputed datasets (Rubin, 1987) depend on the implicit assumption of agreement between the imputer's and the analyst's models (Schafer, 1997: 4.5.4). The imputation model should preserve the associations or relationships among any variables that will be the focus of later analyses (Schafer, 1997). The problem for epidemiology is that it is difficult to anticipate what the focus of future analyses will be since this depends on undetermined scientific and policy relevance (Wadsworth et al., 2003).

1.4 Multiple imputation and analysis in practice

1.4.1 The use of multiple imputation in epidemiological research

An advantage of imputation, shared by multiple imputation (MI), is that it separates the process of dealing with the missing data and the process of analysis. Once completed datasets have been created they can be analysed using any standard complete-data methods.

MI was originally designed for use in complex surveys that are used to create public-use datasets to be shared by many ultimate users, where the data-base constructors and the ultimate users were different people (Rubin, 1996). Creating the multiple datasets is far more computationally demanding than the subsequent analyses. However, the separation of the imputation and the analysis has the advantage that MI can be used by the analyst who does not have the statistical programming expertise required for the imputation step. Despite this practical possibility, MI has not been widely applied in epidemiological research. There are some published studies using MI in epidemiological applications (eg Faris et al., 2002; Arnold and Kronmal, 2003; De Stavola, in press), but until very recently these have been sparse compared to methodological papers. Typical of comments in applied research, Zhou et al. (2001) conclude a review of the use of MI in public health research, *'Although multiple imputation has good statistical properties, it is not yet used extensively.'*

A literature search for multiple imputation in the title of research papers, using the MEDLINE database, produced 39 papers between 1990 and 2002. These were classified according to whether the abstract focussed on the substantive research or on the properties of MI. Of the 22 papers published before 2001, only 5 papers focussed on the substantive result of the research; during the following two years (2001 and 2002) 17 papers were identified, 10 of which focussed on the substantive result (for example Taylor et al., 2002), as did both of the two papers so far retrieved for 2003. For example, in the *Proceedings of the Annual Symposium on Computer Application in Medical Care*, Brand et al. (1994) write: *'Multiple imputation is a statistically*

sound method for handling incomplete data. Application of multiple imputation requires a lot of work and not every user is able to do this.' In 2002, McCleary commented in *Nursing Research* that: *'Accessible, user-friendly computer programs are available to perform multiple imputation for missing data making ad hoc approaches to missing data obsolete.' Nursing Research* is an American journal, as are the publishing journals of the majority of the applied papers, possibly because MI methodology was originally developed in the US. McCleary's (McCleary, 2002) paper refers to Schafer's NORM, freely available from Pennsylvania State University's Department of Statistics website.

In the UK in particular, MI has not been systematically applied to public-use datasets (for example, the 1958 and 1970 British national birth cohorts, which are available from the ESRC data archive), and epidemiologists use many datasets that do not fall into this category (for example datasets collected for a particular purpose, or large datasets which are not publicly available). One reason may be a shortage of statistical expertise or of awareness of the problem of missing data amongst data-base constructors, but in addition, the epidemiologists may be reluctant to accept the results of a statistical process with which they have not been directly involved. The above MEDLINE literature review found that only in 2002 did the substantive research applications in epidemiology outnumber those focussed on methodology. The application of MI in epidemiology in practice has followed on from the release of software procedures in the commonly used commercial software packages (notably SAS in 2002; see Horton and Lipsitz, 2001 for a review of the software packages). The most plausible explanation is that the availability of software has made it feasible for epidemiologists without statistical programming expertise to do their own imputation.

1.4.2 Multiple imputation of alcohol consumption

Although there have been several recent epidemiological publications using MI, applications in the field of alcohol research have been very limited. In 2000 *Addiction* produced a special supplement *State of the Art Methodologies in Alcohol Related Health Sciences Research* in which one paper discusses the treatment of missing data (Figueredo et al., 2000). The authors of this paper echo the problems expressed in general by epidemiologists: *'missing data is a common problem in both cross sectional and longitudinal research. Paradoxically, the problems encountered and the solutions implemented are hardly mentioned outside the statistical literature.'* They discuss a procedure which they call 'Multivariate Imputation' which is used in the context of 'latent variable modelling'. Hawkins et al. (1997) address non-response due to non-initiation of alcohol use in a prospective study of the effects of age at first alcohol use and psychosocial risk factors on subsequent alcohol misuse in male students in the USA. That study used AMOS (**http://www.spss.com/spssbi/amos/**), an add-on to SPSS for structural equuation modelling which has missing data analysis capability. With reference to the more general Schafer techniques for multiple imputation (Schafer, 1997), Figueredo et al. (2000) state that *'not one of these techniques for handling missing data has yet been widely adopted by practicing data analysts.'*

Only one paper was found which uses imputation techniques for missing data specifically on alcohol consumption. This imputes a quantity-frequency (Q-F) measure of alcohol consumption and uses only single imputation methods (Gmel, 2001). Gmel compares the results from imputation using median value replacement, SPSS procedures, and a simple hot deck imputation implemented in Prelis software. Gmel writes *'there are other, more elaborate methods, but they are usually not available to the average survey analyst.'* Gmel is aware of the advantages of MI and Schafer's methods, but does not implement MI because of the lack of available software: *'such implementations would be time consuming and out of reach for a 'normal' survey analyst.'* Schafer's NORM was by then available on his web site as a stand alone windows package and NORM has been applied to handle missing data in the context of analysis of longitudinal data from the1958 birth cohort study (Wiggins et al., 2000). However, as Gmel observes, the use of methods that assume a Normal distribution may be inappropriate for alcohol consumption which is highly skewed (Skog, 1991).

Longford (Longford et al., 2000) applied MI to missing data on alcohol consumption in the diet diaries in the 1946 birth cohort study (the National Survey of Health and Development), programming his own method in the S-Plus language on the basis of specialised statistical knowledge. However, in general, epidemiologists do not have the time or the programming skills to undertake such a task themselves. Hence an aim of this thesis is to examine critically the use of the currently available software to implement ways of dealing with missing data, in the context of epidemiological research.

1.5 Summary

The MRC National Survey of Health and Development (NSHD) is an unusual asset for alcohol research. Not only is it a national survey of the general population, which has followed its members from birth in 1946 and is still doing so, but when study members were aged 43 they completed a seven day diary which included alcohol consumption, a rich source of information on drinking pattern. The problem is that the diary was not fully completed by many of the study members, and this poses the problem of how to deal with the missing data due to such item non-response. The problem arises because the reasons for subjects not completing the diary are unknown and the amount people drink may be related to their failure to complete the diary. However, the strength of the study in relation to this problem is that we do know a great deal about them, including some aspects of their drinking behaviour.

Whereas dealing with missing data has been feasible for analysts with the specialised knowledge and resources to develop their own methods, this has not been possible for epidemiologists until recently, when prepackaged solutions have been made available in some software packages that claim to address the problem. An objective of this thesis was to deal with missing data using existing software, and this involved a critical evaluation of available procedures.

Chapter 2 describes the NSHD, the information on alcohol consumption collected when the survey members were aged 43, and the methods for dealing with missing data that are examined

in this thesis. Chapter 3 justifies the use of the seven-day diet diary as a source of information on alcohol consumption. Chapter 4 demonstrates the danger of ignoring item non-response to alcohol consumption in the diet diary. Chapter 5 investigates the factors associated with non-response and with alcohol consumption in the diet diaries. Chapter 6 uses simulations to develop a multiple imputation method for dealing with item non-response on alcohol consumption in the diary. It takes account of the technical statistical problems which arise with such data and of the characteristics of the data of substantive importance in alcohol research, especially the patterns of alcohol consumption. This entails evaluating the procedures in available software. The sensitivity of the inferences of the prevalence of excessive alcohol consumption to the MAR assumption is evaluated using this simulated data in Section 6.8. Chapter 7 applies the multiple imputation method developed in Chapter 6 to the data on alcohol consumption in the NSHD diet diary. The resulting MI datasets are used to make inferences about the prevalence of excessive alcohol consumption (marginal distribution of an outcome variable with missing values), the role of alcohol consumption in the relationship between birthweight and blood pressure in mid-life (missing values in a potential confounder) and the dependence of blood pressure on alcohol consumption (missing values in the independent variable). It examines the sensitivity of the inferences to dealing with missing data and to the model for imputation. It considers the implication of using the imputations in any subsequent analysis. Finally, Chapter 8 discusses the general implication of the results for dealing with missing data in epidemiological studies of alcohol consumption.

Chapter 2

METHODS

2.1 Introduction

This chapter introduces the MRC National Survey of Health and Development and the information about alcohol consumption collected in 1989 when the respondents were 43 years of age. It discusses the types of missing data in the survey and the methods for dealing with missing data which are tested using simulated data in Chapter 6.

2.2 The MRC National Survey of Health and Development

The MRC National Survey of Health and Development (NSHD) is a follow-up of legitimate, single births to all wives of non-manual and agricultural workers and to one in four wives of manual (but not agricultural) workers in England, Wales or Scotland during the week 3rd–9th March 1946, a sample of 5362 births. A wide range of information on social, psychological, educational, medical and biological characteristics of the study members has been collected on twenty occasions during infancy, childhood and adult life (Wadsworth, 1991; Wadsworth et al., 2003). At the penultimate of these occasions, in 1989, trained nurses interviewed the study members when they were 43 years old. At this time, 3262 (85%) of the 3854 with whom contact was attempted were interviewed; 4 (0.1%) had died, 11 (0.3%) were living abroad, 106 (2.7%) were permanent refusals, 195 (5.1%) temporarily refused because of personal or family problems and 276 (7.2%) could not be traced. Of the 1508 of the original birth sample whom there was no attempt to contact, 361 (24%) had died, 607 (40%) were living abroad and 540 (36%) had permanently refused to take part at a previous contact (Wadsworth et al., 1992). Excluding study members who were living abroad, whom the survey did not intend to represent, and those who had died, 74.5% (3262/4379) of those in the birth cohort who were still alive and resident in England, Wales or Scotland were interviewed at the age of 43.

2.3 Measures of alcohol consumption and drink problems collected at age 43

The information about alcohol was collected at the end of the 1989 interview in two distinct ways: using a self-completion questionnaire and using a diet diary.

The self-completion questionnaire (Appendix 1), contained questions about alcohol consumption during the last seven days (*weekly recall*) and also the CAGE questionnaire (Ewing, 1984).

2.3.1 CAGE

The four CAGE questions, each with yes/no response options, are as follows:

Have you ever felt you ought to *C*ut down on your drinking? (Do not include dieting.)

Have people ever *A*nnoyed you by criticising your drinking?

Have you ever felt bad or *G*uilty about your drinking?

Have you ever had a drink first thing in the morning to steady your nerves or to get rid of a hangover? (*E*ye-opener)

The questions were asked of lifetime experience of such problems, and where the answer to a question was 'yes', study members were asked whether they had experienced this 'in the last year'. The CAGE score is defined as the number of affirmative answers to these questions, ranging from 0 to 4; there are two scores, one relating to lifetime experience of problems ('CAGE EVER') and one relating to problems in the last year ('CAGE LAST YEAR'). The latter score is used in the analyses in this thesis as it is more relevant to current drinking levels than the former.

2.3.2 Weekly Recall

Consumption of alcohol was based on the responses to the question 'In the last seven days how many of the following drinks have you had?' Three categories of drink were differentiated: spirits (measures of spirits or liqueurs); wine (glasses of wine, sherry, martini or port) and beer (half pints of beer, lager, cider or stout). The number of drinks reported were thus approximately equivalent to the number of Units of alcohol, the commonly used UK measure of alcohol.[1]

2.3.3 Seven-Day Diet Diary

The seven-day diet diary was used to record all food and drink, including alcohol, consumed during each day of one week. The diary was begun at the end of the interview when the research nurse interviewer recorded all food and drink consumed in the previous two days, demonstrating the method and the detail required. The study member was then asked to keep the diary for the subsequent five days and to return it by post in the pre-paid envelope supplied. A carbon copy of the first two days was retained by the nurse, so that at least two days of dietary information was available from those who were interviewed. Thus the first two days of the diary were recorded retrospectively by the nurse who prompted the study member for information; the subsequent five days were completed by the respondents themselves.

The diary (Appendix 2) comprised daily sheets, each identified by date and the day of the week, providing spaces in which to record meals, including alcoholic and non-alcoholic beverages and between-meal snacks, ending with a reminder section to record any other snacks or drinks not

[1] 'A Unit is roughly equivalent to half a pint (290 ml) or ordinary (4 per cent) beer or lager, 1 pub measure (24 ml) of spirit, 1 glass (50 ml) of sherry or fortified wine or 1 glass (125 ml) of wine.' (Faculty of Public Health Medicine, 1996).

previously recorded. This reminder section included specific prompts for beer, wine, sherry or spirits besides such items as sweets, tea, and other common items of food which may not have been taken with a meal. The layout of the sheets was structured in three columns headed: 'food/drink', 'description and preparation' and 'amount'. The weight and nutritional composition of all food and drink recorded in the diaries was derived by the MRC Dunn Nutrition Unit, Cambridge, using a computerised system, DIDO (Diet In Data Out), together with a suite of programs for nutritional analysis (Price et al., 1995). The alcoholic content of the drinks was converted to grams of alcohol per 100ml. This conversion was based on a study of the average alcohol content of 29 types of beers, ciders, wines, liqueurs and spirits derived from samples of each type (Paul and Southgate, 1978). The total alcohol consumed per day (in grams) was calculated from the quantities and types of drink reported. The data used in the analysis below consists of 7 items per subject, each item being the alcohol consumption on each day of the diary. Where appropriate, grams of alcohol were converted to Units by dividing by 7.9 (Royal College of Psychiatrists, 1986).

2.4 The distribution of alcohol consumption

The distribution of daily alcohol consumption is illustrated by the reported alcohol consumption on Saturday of the diary week (Figure 2.1).

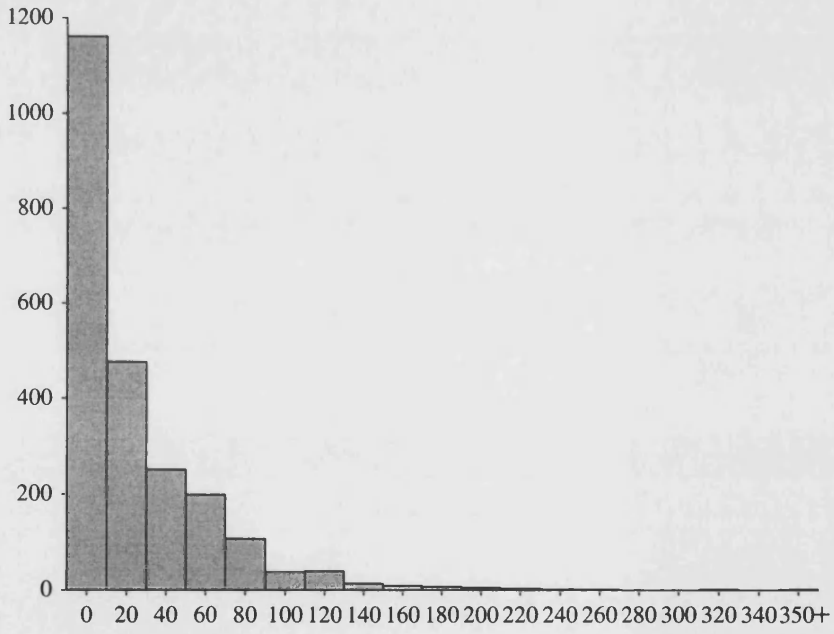


Figure 2.1: Histogram of Alcohol Consumption (grams) on Saturday of the diary week by 2316 respondents who completed the diary for that day.

The distribution is positively skewed, most people drinking nothing or a small amount, and few people drinking large quantities of alcohol. Positively skewed data is often transformed using a logarithmic transformation (see e.g. Altman, 1991). This transformation changes multiplicative difference into additive ones; in other words, it tends to spread out low values and compress

high values, giving a more 'Normal' shape to the distribution. Where the variable has zero values, however, a logarithmic transformation requires the addition of a constant (such as 1), since the logarithm of zero is not defined. Figure 2.2 shows the result of applying such a transformation to the alcohol consumption on Saturdays shown in Figure 2.1.

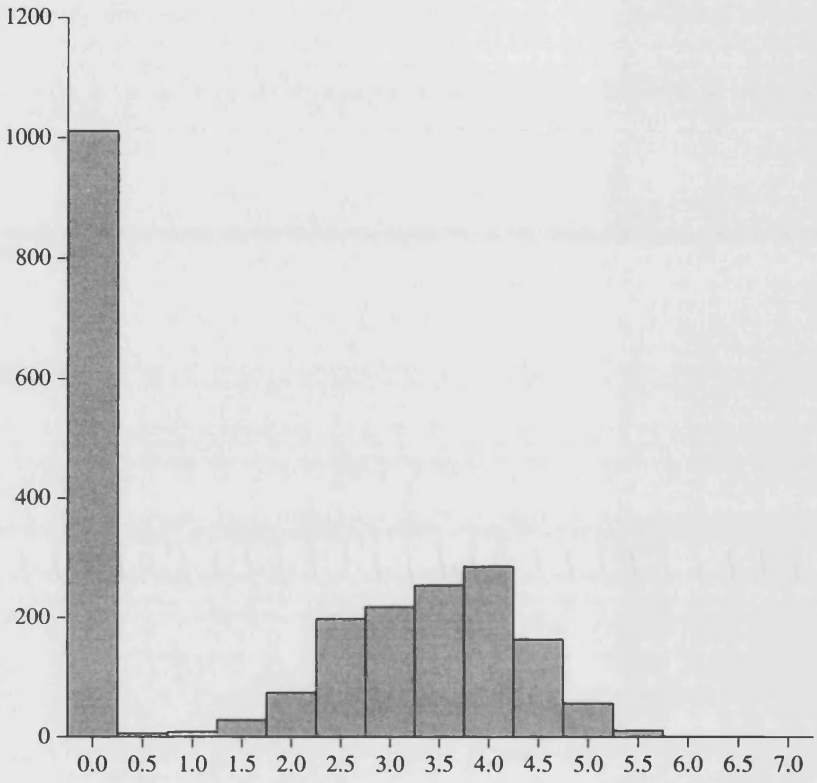


Figure 2.2: Histogram of the quantity log(1 + *ALC*) where *ALC* is the Alcohol Consumption (grams) on Saturday of the diary week by 2316 respondents who completed the diary for that day.

The result of applying this transformation is a distribution in two parts: a set of zero values (for those who did not drink on the Saturday) and a separate continuous distribution which is approximately Normal (for those who drank something on the Saturday, the logged positive amounts). A variable with this kind of distribution, characterised by a proportion of responses equal to a single value (often zero) and a continuous distribution amongst the remaining responses, is called semicontinuous (Olsen & Schafer, 1998). It arises here because a high proportion of people drink nothing at all. This poses a problem when using parametric methods that assume the data to be Normally distributed, since they may produce erroneous estimates. Semicontinuous variables occur in various fields of research. Besides alcohol consumption, examples include tobacco consumption, blood alcohol content in motorists, annual household expenditure on specific goods and income from specific sources (Olsen & Schafer, 1998).

2.5 Missing data in the NSHD

2.5.1 Formal definitions of missing data mechanisms

First some notation is introduced. Let Y denote a variable with missing values. We partition Y into two parts:

Y_{obs} containing the observed data, and

Y_{miss} containing the missing part of Y.

X denotes a set of covariates, $x_1, x_2, \ldots, x_p$ which we assume are fully observed.

R is the response indicator (equivalent to a missing data indicator $M = 1 - R$), defined as

$R_i = 1$ when Y_i is observed,

$R_i = 0$ when Y_i is not observed.

There are three types of missing data mechanisms (Rubin, 1987):

MCAR — Missing completely at random

The probability that a response is missing depends on neither the response variable nor the covariates.

$$P(R \mid X, Y) = P(R) \tag{1}$$

This is a very strong assumption, that observation of data values does not depend on any observed background variables or on the observed or unobserved values of the outcome variable. The unobserved responses are a random sample of the observed data. For example, the probability that the respondent does not complete a diary day depends neither on the amount they drank on that day, nor on any other recorded variable such as the amount they drunk on the previous day, nor on the amount they recalled drinking in the seven days prior to the interview.

MAR — Missing at random

The probability that a response is missing depends on the value of the covariates and on the observed values of the variable Y.

$$P(R \mid X, Y) = P(R \mid X, Y_{obs}, Y_{miss}) = P(R \mid X, Y_{obs}) \tag{2}$$

For example, the probability that the respondent does not complete the diary day depends on, say, their gender, the amount they drank on the previous day, or the amount they recalled drinking in the seven days prior to the interview, and so on. The value of an entry missing in the diary may depend on the amount drunk, but only through the relationship of the observed values with the covariates.

MNAR — Missing not at random or non-ignorable missingness

The probability that a response is missing depends on the value of the covariates and on both the observed and unobserved values of the response variable.

$$P(R \mid X, Y) = P(R \mid X, Y_{obs}, Y_{miss}) \tag{3}$$

For example, the chance of a respondent not completing a diary day is greater if they consume a large quantity of alcohol on that day, even after allowing for the association of consumption with observed values of other variables.

2.5.2 Types of missing data in the MRC NSHD

There are three types of missing data in the information collected by the NSHD survey at follow up interviews. These are:

A. Missing by design, due to the stratified sample design. The sample of births was stratified according to the father's occupation. All births to wives of non-manual and agricultural workers were included in the sample, but only 1 in 4 of births to wives of manual (but not agricultural) workers.

B. Missing due to the study member not having been interviewed (case non-response or attrition). We can define two subsets of these:

   B.1 Missing because dead or emigrated. Those who emigrated were not included in the sample whilst resident abroad, although attempts were made to keep in touch with them so that they could be included if they resumed residence in Britain.

   B.2 Missing for other reasons — for example the subject cannot be traced, refuses to be interviewed, or is unavailable for interview.

C. Missing due to non-response to the interview item or question (item non-response).

The methods we use to deal with missing data should take account of knowledge about the process of missingness.

For cases which are missing by design (A) or because the study member has died or emigrated (B.1) the mechanism of missingness is known, and therefore they are simpler to deal with. We know that the cases missing by design are missing at random (MAR) within strata defined by the stratifying variable (based on the occupation of the subject's father). Those who were dead could not be in the sample, and those who had emigrated were not intended to be.

Whether we are concerned with the other type of case non-response (B.2) depends on the level of inference or the definition of the population which we want to make inferences about. If we wish to make inferences about native born subjects who are still alive and living in England, Scotland and Wales, then case non-response by those who could not be traced or refused to be interviewed in 1989 (B.2) has to be dealt with in addition to item non-response. If, however, we

are concerned only to make inferences to those who are represented by the sample interviewed in 1989, then we need to deal with item non-response only (C).

For both of these types of missing data (B.2 and C), the mechanism of missingness is essentially unknown. We can only make assumptions about the mechanism of missingness and examine the sensitivity of our inferences to those assumptions. However, the reasonableness of the assumptions may also be assessed in the light of our knowledge (external and internal to the data being analysed) of the variable.

2.6 Methods for dealing with missing data due to case non-reponse

The method of dealing with missing data depends on our knowledge, or assumptions, about the missing data mechanism. If the sample we have for analysis is a simple random sample of this population then we know that the data is missing completely at random (MCAR). In practice, the epidemiologist using survey data has at his disposal a sample of people in which not everyone who was designed to take part does so, or in which those who do take part do not provide all the information intended to be collected, and in general the precise reasons for this non-response are not known. When the missingness is uncontrolled and unintended by the sample design, as for B.2 and C above (Section 2.5.2), the mechanism of missing data is generally unknown, and can only be conjectured. Under the assumption of MCAR complete-case analysis gives unbiased estimates, but MCAR is very unlikely. It is easily disproved by finding a single variable which is related to missingness.

Under the MAR assumption the information contained in the incomplete records can be used through the relationship of the observed covariates with the observed items of the variable with missing values. The observed data can be exploited by methods based on modelling the variable with missing values using the MAR assumption. The assumption of MAR cannot be tested empirically. Indeed, in the example of alcohol consumption it is plausible that the amounts people drink may be related to their failure to complete the diary; if this is the case, then the diary items are MNAR. However, the plausibility of the MAR assumption is relative to the data used. For example, suppose that alcohol consumption is missing at random conditional on smoking status: that smokers drink more than non-smokers and smokers are more likely not to complete their diaries. In the absence of data on smoking status, alcohol consumption will be MNAR; but if smoking is observed then alcohol consumption is MAR conditional on that information. The MAR assumption fails to hold only if non-response depends on the missing values after conditioning on the available information. It follows that the MAR assumption is more plausible the richer the set of conditioning variables included in the model for missing data (Rubin, 1987).

2.6.1 Missing by design

By definition, we have no information about Y (alcohol consumption) or any covariate X in the cases not selected to be in the sample, but we know the mechanism of missingness so that we know that Y is MAR conditional on the strata membership. The simplest solution to the data missing by sampling design is to employ weighting.

The birth cohort consisted of all legitimate single births during the week 3–9 March 1946. The cohort was divided into two strata according to the occupation of the father. The sample included all single legitimate births to wives of non-manual and agricultural workers (stratum *a*), $n_a = 2992$ cases in all, and a simple random sample of 1 in 4 from such births to wives of manual workers (stratum *b*), giving $n_b = 2370$ cases in the sample. There are therefore $n = n_a + n_b = 5362$ cases in the overall sample, of which 2992 represent 100 per cent of their cohort stratum and 2370 represent 25 per cent of their cohort stratum.

By definition, we do not know the values of Y (alcohol consumption), or any covariate X, in the cases not selected to be in the sample, but we know the mechanism of missingness so we know that Y is MAR conditional on stratum.

When the objective is to estimate a quantity for the whole population, the simplest solution is to combine the estimates of this quantity for the different strata, using weights which reflect the ratios of numbers in population to numbers in sample according to stratum.

Let Q be a quantity which is estimated from stratum a of the sample by $\hat{Q}_a$, with variance (SE squared) V_a, and from stratum b by $\hat{Q}_b$ with variance V_b. For the whole cohort, the quantity Q is estimated as

$$\hat{Q} = \frac{n_a\hat{Q}_a + 4n_b\hat{Q}_b}{n_a + 4n_b} \tag{1}$$

Let weights w_a and w_b be defined as

$$w_a = \frac{n_a}{n_a + 4n_b}, \qquad w_b = \frac{4n_b}{n_a + 4n_b} \tag{2}$$

Then

$$\hat{Q} = w_a\hat{Q}_a + w_b\hat{Q}_b \tag{3}$$

i.e. a weighted mean of $\hat{Q}_a$ and $\hat{Q}_b$ with weights w_a and w_b. $\hat{Q}_a$ and $\hat{Q}_b$ are uncorrelated because stratum b in the sample is a simple random sample from stratum b in the cohort, therefore the variance of $\hat{Q}$ is

$$V = w_a^2V_a + w_b^2V_b = \frac{n_a^2V_a + 16n_b^2V_b}{(n_a + 4n_b)^2} \tag{4}$$

2.6.2 Subjects not interviewed

We have no concurrent (cross-sectional) information about the alcohol consumption at age 43 of members who were not interviewed at this age. Of course, we may have a large amount of information about them from previous occasions when they have been interviewed. Attrition tends to increase with the lifetime of the cohort. Thus variables measured in earlier life tend to be more fully observed, but the association of such variables with alcohol consumption at age 43 is likely to be weaker. Variables recorded at earlier interviews provided little information about alcohol consumption at age 43. In this study inferences are made about the population well represented by the subjects interviewed in 1989. The issue of case non-response by those not interviewed is not dealt with.

2.7 Methods for dealing with missing data due to item non-reponse

Dealing with item non-response is the principal focus of this thesis, although the methods discussed here are not exclusively used for item non-response, but can be used for any missing data where the missingness is unintended and therefore the mechanism of missing data is generally unknown. First some general terms and characteristics of methods are introduced.

2.7.1 Monotone missing-data pattern

A monotone missing-data pattern (Little, 1995) occurs when the variables with missing values can be ordered from left to right, such that whenever any variable is observed then every variable to the left of it is also observed. Let the variables with missing values be Y_i; the Y_i are ordered such that for any case where the response indicator $R_j = 1$ (Y_j observed), we have $R_i = 1$ for all $i < j$. For example: if, when anyone fails to complete a day in a diary they also fail to complete any subsequent day, then the diary has a monotone missing pattern. If the variables have a monotone pattern, then the cases can be ordered into a 'staircase' pattern as shown in Figure 2.3.

		Variables				
		1	2	3	4	5
Cases	1	•	?	?	?	?
	2	•	•	?	?	?
	3	•	•	•	?	?
	4	•	•	•	?	?
	5	•	•	•	•	•
	...					

Staircase pattern

Figure 2.3: Illustrating monotone missing-data patterns

• Observed data ? Missing data

2.7.2 A general taxonomy of methods

Complete Cases only

This method uses only the cases which have all variables observed, discarding all cases that have any missing values. This is often described as Listwise Deletion (LD) in software packages. Most statistical procedures in standard software packages employ this method, and where there are alternatives available it is usually the default, and it is adopted without warning the user.

Imputation

Imputation means filling in the missing values. The completed data can then analysed using standard methods. A ***deterministic imputation*** is completely determined by the observed data; in a ***stochastic imputation*** the values imputed incorporate a random element. The advantage of a stochastic imputation is that it can be arranged so that it preserves the sampling variability in the variable with missing values (Y), whereas a deterministic imputation cannot.

If each missing value is replaced with a ***single imputation*** and the completed dataset is analysed by standard methods, the imputed values are treated as if they were observed. The uncertainty about the missing values is not reflected in the analysis. An imputation which is deterministic can only be a single imputation, but a single imputation may not necessarily be deterministic. ***Multiple imputation*** (MI) imputes a number (m) of plausible values for each missing value. These are used to make up m completed datasets which are identical for any observed values and have (in general) different imputed values in place of the missing values.

Each of these m datasets can then be analysed using any standard method to derive the required estimates. The estimates and sampling variances from each of the m datasets are combined by averaging. The additional uncertainty due to the missing values can be represented by the variance between the m estimates. The details of how the estimates from MI are derived is explained below in section 2.9.

When a variable Y_i has missing values we have essentially two sorts of information to draw on. We may use the information we have about the response indicator R, or the information about Y_i itself. An approach which models R is called Propensity Score (see 2.8.4.1 below). This has the advantage of being non-parametric since the dependent variable is binary. The problem with this approach is that information about R may not be relevant to the value of Y_i. For example, knowing that someone does not report their drinking because they are illiterate does not inform us about what they drink unless those who are illiterate have a distinct pattern of drinking.

In order to make a plausible imputation of a missing value (a good guess), say of what a person drinks, it is sensible to take into account what we do know about that person which could have a bearing on their drinking. For example it is well known that women drink less than men so if we know the person is a woman we would guess that they drank a smaller amount than if they were a man. We might use the mean observed alcohol consumption of women as an imputation

for a woman, and similarly for men. This is known as ***mean value replacement*** (MVR). However, MVR is deterministic and distorts the distribution of the variable being imputed, because there is a concentration of values at the mean point. It also distorts the covariances and correlations with other variables. Alternatively one can replace missing values with a randomly drawn observed value from a matching group. For example the imputated value of alcohol consumption for a woman could be a random draw from the observed alcohol consumption for women. This is known as ***Hot Deck*** imputation. This is a stochastic imputation method that does not distort the distribution, but it does distort the correlations with other variables.

The background variables that we use can be called covariates (X) for the imputation of the variable(s) with missing values (Y). We use as covariates variables that are related to the variable to be imputed (Y). The more covariates we can take into account, the more plausible is our imputed value. For example, a woman with missing alcohol consumption who smokes is likely to drink more than a woman who does not smoke. With methods like MVR and Hot Deck the covariates used are categorical: we divide the data into groups according to the values of those categorical covariates. The more categorical covariates we use the greater the number of different groupings (or cross-classifications) of the data and the smaller the number of cases in each group. This poses a problem because the smaller the number of cases we have in a group, the greater the uncertainty about the values in the group.

To avoid dividing the data into groups we can base our imputations on a model for the distribution of the missing variable, for example by regression. Using regression, we take the variable with missing values (Y_i) to be the dependent variable in a regression model, and we model the dependence of Y_i on the chosen covariates (which may be X and $Y_j, j \neq i$). The imputed values for Y_i are randomly drawn from the distribution implied by the regression model conditional on the observed independent variables in that model (X or the $Y_j, j \neq i$). Standard regression uses only cases with complete data, so this will use all the observed data only for those cases with complete data.

A method which can take into account all of the observed data in the partially completed records (of X or of Y) will be most efficient. Regression can be adapted for missing data by estimating the coefficients using pairwise available statistics (i.e. for each pair of variables we use cases that have both variables observed to estimate the covariance). This approach is used, for example, by SPSS Regression which is fully described below (Section 2.8.3.1). This method fails to use all of the observed data because cases with variables which are pairwise missing cannot contribute to the estimated covariance. Also it does not take into account all the observed data at the same time. A separate regression must be fitted for each missing variable in turn, and each one will include a different set of cases.

Another way to make more use of the observed data with regression is possible when the data (X, Y) have a monotone missing pattern (Section 2.7.1). When the data are monotone missing, they can be arranged in a staircase pattern and it is possible to impute the variables with

missing values step by step starting with the variable with the least missing values. Separate regressions are fitted for each Y_j $(Y_j | X, Y_1, Y_2, \ldots, Y_i)$ where $i < j$, using the observed values of all the Y_i such that $i < j$, and previously imputed values of these variables (The covariates X are here assumed to be completely observed). This approach has the disadvantage that it does not take into account the relationships between the Y_k and Y_j where $k > j$. Another disadvantage occurs if the data is not monotone missing. To make use of this approach when the data is non-monotone missing, the variables that destroy the monotone structure (including any covariates with missing values) must be imputed first using ad hoc procedures. This approach is used by the SOLAS software for multiple imputation which is described below (Section 2.8.4).

Each of these methods makes use of all the observed data, but they use a series of separate regressions, each of which uses a different subset of the observed values. Hence they cannot take into account all of the observed data simultaneously. This has two disadvantages. Firstly, they are not fully efficient. Secondly, they cannot take full account of the interrelationships between all the variables in the set (X, Y).

These disadvantages are overcome in methods which use the whole dataset (X, Y), dealing simultaneously with all the variables, not distinguishing between X and Y in the way they are treated. These define a joint distribution for the partially missing data (X, Y), for example multivariate Normal, and generally use maximum likelihood based on the full data to estimate the parameters for a distribution implied by the model. This is most easily achieved computationally using the EM algorithm.

Proper imputation

It is important for valid inference using multiple imputation that the imputation procedure should be ***proper***. When imputations are made, the imputed values should have a random component which reflects two sources of uncertainty. First, the model itself has only been estimated (using the available information in the observed data), and so its parameters are uncertain. Second, the values of the missing data are uncertain because they could be any values sampled from the distribution specified by the model. So, in practice, proper imputation proceeds by sampling random parameter values for the model, and then sampling imputed values according to this random model, repeating both steps independently for each imputation.

This ensures that the variation between imputed values in different imputations corresponds to the total uncertainty about the underlying values of the missing data, due to the uncertainty about the model, and the randomness in the variables given the model.

2.7.3 The EM Algorithm

The EM Algorithm (Dempster, Laird and Rubin, 1977) was devised for maximum likelihood estimation from data with incomplete information. It implements an iterative procedure for estimating the parameters in an assumed distribution, which would be estimated by a straight-

forward calculation if the data were complete (for example, assuming a multivariate Normal distribution, the means and the covariance matrix).

Each iteration consists of two steps, an 'E' step and an 'M' step. The E step computes the expected values of the sufficient statistics for the parameters, initially using a starting value for the parameters, and subsequently using the parameter estimates derived from the M step. The M step uses the expected values of the sufficient statistics from the E step to re-estimate the parameters by maximum likelihood estimation; these are then used in the following E step. The steps are repeated until the parameters produced at the M step are sufficiently close to those obtained at the previous M step, when we say that convergence has been achieved.

The following procedures are evaluated using simulated data in Chapter 6:

Complete cases only
Mean value replacement
SPSS Regression
SPSS EM
SOLAS Propensity Score
SOLAS Discriminant Method
SOLAS Predictive Model Based Method
Schafer's CAT (or S-Plus Loglinear)
Schafer's NORM (or S-Plus Gaussian)
Schafer's MIX (or S-Plus Conditional Gaussian model)

They are described in detail in the next section.

2.8 Procedures for dealing with item non-reponse

2.8.1 Complete cases only

This is described above (Section 2.7.2)

2.8.2 Mean value replacement

Missing values are replaced by the mean of the observed values of cases in the same group.

2.8.3 SPSS Missing Value Analysis (MVA)

SPSS v 7.5 (SPSS, 1991), and subsequent releases of the software, offer two options for imputing missing variables, called Regression and EM, in their MVA procedures. Both of these are applied to continuous variables only: the variable to be imputed and any variables used as covariates must be declared as quantitative in the procedure. Both procedures are designed to produce single imputations only.

2.8.3.1 SPSS Regression

SPSS Regression (SPSS, 1997) fits a least squares linear regression model for each variable with missing values in turn with all other variables selected for the model as covariates. Essentially, the SPSS Regression approach estimates means as the mean of all available data for each

variable, and covariances as the covariance between pairs of variables calculated over cases where data for both variables in a pair are available. These estimates are used in the standard equation which predicts the expected value of one variable in a multivariate Normal conditional on the values of other variables. Then an imputed value can be generated from this formula conditional on the values of the observed variables. The user can choose to add a random component to the imputation so that the imputation is stochastic.

By default the method fits the least squares linear regression model for each variable with missing values in turn, with all other variables that are declared as quantitative used as covariates. The user can select which of the quantitative variables are to be used as predicted and which as predictors. The user can also specify the maximum number n of predictor variables to use (in which case the n best from forward stepwise selection will be used), and a threshold based on an F test for significance of a predictor variable for entry of a variable into the regression (which may have the effect that fewer variables are used). Once the missing values have been estimated, by default they are perturbed by adding a randomly chosen residual from the linear regression or, at the user's option, a random value drawn from a Normal or a t distribution, or no perturbation is applied at all. The following gives the details of the procedure, based on SPSS (1997).

To impute a missing value x_{ij} — for the ith case of the jth variable — the method generates a predicted value

$$\hat{x}_{ij} = \hat{\beta}_{0,ij} + \sum_{l} \hat{\beta}_{l,ij} x_{il} + \varepsilon_{ij}$$

where in the summation l takes values along row i such that x_{il} is not missing (and $l \neq j$), and the coefficients $\hat{\beta}_{l,ij}$ have been estimated from the data as described below. Note that there is a different set of regression coefficients for each (i, j), i.e. for each missing value in the data. The term ε_{ij} is an optional random error term.

The description given in the SPSS documentation of how the regression coefficients are calculated is limited to the following (quoted from SPSS, 1991):

> $[\hat{\beta}_{0,ij}, \hat{\beta}_{l,ij}]$ is computed from Diag($\bar{\mathbf{X}}^P$) = $[\bar{x}_j^P]$
> and by pivoting on the "best" "q" of the J_1 diagonals of $\mathbf{C}^P$.
> "best" is forward stepwise selected.

This is difficult to interpret, and the following seems to be a correct explanation of it; the notation has been modified, for clarity.

The statistics from which the regression coefficients are derived are what SPSS calls "pairwise statistics".

They are

- The variable-wise sample means:
  For each variable x_j the mean $\bar{x}_j$ of the non-missing values of that variable;
- The pair-wise sample means:
  For each pair of variables x_j and x_k, the mean $\bar{x}_{j|k}$ of x_j taken over all cases where neither x_j nor x_k has missing values (i.e. both variables observed), and the mean $\bar{x}_{k|j}$ of x_k taken over all cases where both variables are observed. When $j = k$, it is the same as the variable-wise sample mean above and corresponds to SPSS's expression $\mathrm{Diag}(\bar{\mathbf{X}}^P) = [\,\bar{x}_j^P\,]$ above, since the pairwise means can be arranged in a matrix $\bar{\mathbf{X}}^P$ where the element in row j and column k is $\bar{x}_{k|j}$.
- The sample covariance matrix $\mathbf{C}^P$ calculated from the pairwise complete statistics. The (j,k)th element c_{jk} of $\mathbf{C}^P$ is given by:

$$c_{jk} = \frac{1}{n_{jk}-1}\sum_i (x_{ij} - \bar{x}_{j|k})(x_{ik} - \bar{x}_{k|j})$$

  where the summation is over cases i where both variables x_j and x_k are observed, and n_{jk} is the number of these cases.

In the standard regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\mathbf{y}$ denotes a column of the n values of the dependent variable, the matrix $\mathbf{X}$ has a row corresponding to each value in $\mathbf{y}$ consisting of the values of the independent variables $x_1, x_2, \ldots, x_p$, and $\boldsymbol{\beta}$ denotes a column of the regression coefficients, the ordinary least-squares estimate of the coefficients $\boldsymbol{\beta}$ is given by the matrix equation

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

If all the variables are 'centred at their means' (i.e. have their means subtracted throughout), then for n observations $\mathbf{X}^{\mathrm{T}}\mathbf{X}/(n-1)$ is the sample covariance matrix $\mathbf{C}$ of the $\mathbf{X}$ variables, and $\mathbf{X}^{\mathrm{T}}\mathbf{y}/(n-1)$ gives the sample covariances $\mathbf{C}_{\mathbf{y}}$ between $\mathbf{y}$ and the $\mathbf{X}$ variables. So the above equation can be written

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = (\mathbf{X}^{\mathrm{T}}\mathbf{X}/(n-1))^{-1}(\mathbf{X}^{\mathrm{T}}\mathbf{y}/(n-1)) = \mathbf{C}^{-1}\mathbf{C}_{\mathbf{y}}$$

Hence, using centred variables, the regression coefficients of one variable on other variables can be computed from their sample covariances. Once this is done, the variables can be 'de-centred' by adding back their means which, when multiplied by the regression coefficients and collected up, give the intercept term.

It remains to interpret 'pivoting on the "best" "q" of the J_1 diagonals of $\mathbf{C}^P$'. The SPSS Regression method allows the user to specify a maximum number of variables to include in the regression, and also includes a default threshold value of 4.0 for the F test for inclusion of a variable in the regression by forward stepwise selection. Hence, to impute a particular missing value x_{ij} in the data, note that this is a value of the jth variable; it corresponds to the variable denoted by **y** above. Hence the relevant elements of the covariance matrix $\mathbf{C}^P$ are

- The subset which does not include the jth row and jth column and which includes the rows and columns corresponding to the other variables currently considered for inclusion, which make up the covariance matrix **C** in the above equations;
- The elements in the jth row (or column) of $\mathbf{C}^P$ which make up the covariances $\mathbf{C}_y$ between **y** (i.e. x_{ij}) and the other variables.

Now the above method of calculation can be applied to give the regression coefficients and, if appropriate, a further variable can be considered for inclusion. If the analysis of variance for inclusion of this variable gives an F ratio exceeding the threshold, then it will be included, and so on until either no further variables satisfy the F criterion for inclusion, or the designated maximum number of variables have been included.

The random error term ε_{ij} may be chosen from

i residual of a randomly selected complete case

ii random Normal deivate, scaled by the standard error of the estimate

iii random t deviate, scaled by the standard error of the estimate and with degrees of freedom specified by the user

iv no random error term

2.8.3.2 SPSS EM

SPSS EM (SPSS, 1997) uses the EM algorithm (Section 2.7.3) for maximum-likelihood estimation of the parameters (the means and the covariance matrix) in a multivariate Normal distribution. In SPSS EM, the E step computes the expected values of the individual missing data, as well as of the sufficient statistics, initially using the parameter estimates as computed by the SPSS Regression procedure, and subsequently those derived from the M step.

The imputed values generated by the SPSS EM procedure are taken to be the expected values of the missing data obtained in the final E step. It is not possible in SPSS EM to vary these values in any way to reflect the uncertainty in these imputations. The user can specify the assumed distribution of the data. By default the distribution is assumed to be multivariate Normal but the user can specify alternatively a multivariate t with chosen degrees of freedom, or a mixture of two Normal distributions with chosen proportions and variance ratio.

2.8.4 SOLAS procedures for multiple imputation

SOLAS (V 2.0 and subsequent releases) includes two distinct types of procedure for multiple imputation (MI): model based procedures, which model the variable to be imputed (Y), and a procedure which models the missingness of the variable to be imputed (R). The latter is called 'Propensity Score'. Propensity Score was the only MI procedure offered in SOLAS version 1.0 (Statistical Solutions, 1997). There are two model based procedures: the 'Discriminant Method' for imputing categorical variables and the 'Predictive Model Based Method' for imputing continuous variables. These procedures are described individually below but we first describe the general principles on which all the SOLAS procedures work.

Each of the SOLAS procedures requires the covariates (X) to be completed first. The user can chose to handle missing covariates in one of three ways: they can be imputed using Hot Deck (Section 2.7.2), or a missingness indicator for the covariate can be included in the regression pool, or cases that have missing values in this covariate can be excluded from the analysis.

If Hot Deck is specified SOLAS replaces a respondent's missing value with a value randomly selected from matching respondents, i.e. those who have the same value on a covariate chosen by the user.

The SOLAS MI procedures start by sorting the data into a structure as near as possible to a monotone missing pattern. It then imputes the missing values which destroy the monotone structure by using cases with the same pattern of missing data but observed on the variable to be imputed, until a monotone structure is achieved. It is possible to specify a different set of covariates for imputing the non-montone and montone missing values in the imputation variable (Y).

With the data in a monotone structure, SOLAS proceeds in a step by step fashion, imputing each variable one at a time using separate regressions, starting with the least missing variable, and proceeding from left to right to the variable with the most missing values. The regression uses the covariates and also any previously imputed variables, as specified by the user, as independent variables in the regression. Because separate regressions are used it is possible to specify an entirely different set of covariates (X) for each imputation variable (Y).

2.8.4.1 SOLAS Propensity Score

The SOLAS Propensity Score procedure models the probability that a variable is missing (the propensity score) and draws imputations from the observed values of the variable for cases which have a similar propensity to be missing (called the donor pool) (Lavori et al., 1995). It uses the response indicator R_i (see below) as the dependent variable in a logistic regression. The following description is derived from the SOLAS user reference manuals for version 1.0 (Statistical Solutions, 1997) and version 2, especially Appendix E, (Statistical Solutions, 1999).

For a given variable y with missing values, the procedure is as follows:

1 A 'response indicator' variable R_y is created, which takes the value 1 for a case where y is observed, and takes the value 0 for cases where y is missing.

2 SOLAS allows selection of which variables shall be considered as covariates for a given variable y in imputing missing values. Let these be $x_1, x_2, \ldots, x_k$. A logistic regression

$$T = \log\frac{P}{(1-P)} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

of $P = \text{Prob}(R_y = 1)$ on $x_1, x_2, \ldots, x_k$ is estimated. This is equivalent to

$$P = \frac{e^T}{1+e^T} = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}{1+\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}$$

and then $1 - P$ is the conditional probability of missingness, given the values of the covariates.

3 This gives estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ of the βs, and hence for each case i an estimated value of T:

$$T_i = \hat{\beta}_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i}$$

and hence also of the probability

$$P_i = \frac{\exp(T_i)}{1+\exp(T_i)}$$

According to the documentation for SOLAS version 1.0, "the propensity score is the conditional probability of missingness. given the vector of observed covariates." On this basis, the estimated propensity score for case i would be the value of $1 - P_i$ as above.

However, the documentation for SOLAS version 2.0 states that to each case "a propensity score is assigned which is equal to $X_i^T b$"; in other words, the value of T_i as above. However, it is later pointed out that in the use which is made of the score it is equivalent whether $1 - P_i$ is used or T_i, and "the propensity scores [T_i] are used rather than these estimated probabilities for reasons of numerical stability." In the following, "propensity score" refers to T_i.

4 The cases in the dataset are sorted in increasing order of propensity score.

5 For each case of missing data for y, a subset of cases with observed values of y is selected which will have similar propensity scores to the score for the case with missing y. SOLAS allows this do be done in any chosen one of the following ways:

   a Divide the propensity scores into a number of quantile subsets (the default is 5 quintiles), and choose the subset in which the propensity score for the missing case falls.

b Choose, for a given number C, the C cases whose propensity scores are closest to that of the missing case.

c Choose, for a given percentage D, the $D\%$ of cases whose propensity scores are closest to the missing case.

d Each of the above (a), (b) or (c) can be augmented with a chosen "refinement variable" w; the subset arising as in (a), (b) or (c) is further reduced to a specified number consisting of those which are closest to the missing case in their value of w.

6 The result of stage 5 is a 'donor pool' of cases with observed data on y, chosen so as to have similar propensity score to the missing case.

7 Let there be K cases in the donor pool. Next, a random sample S of size K is drawn with replacement from the donor pool.

8 Let there be L cases with missing y associated with a given donor pool. Next, a random sample of size L is drawn, with replacement, from the random sample S, and the y values in this second sample are used to fill in the missing values.

2.8.4.2 SOLAS model based procedures

When the user opts for 'Predictive Model Based Method', the software automatically uses what the manual calls 'Discriminant Method' (Section 2.8.4.2.1 below) when the variable to be imputed is nominal, and what the manual calls 'Predictive Model Based Method' (Section 2.8.4.2.2 below) when the variable to be imputed is continuous or ordinal categorical. To avoid confusion here the name 'Predictive Model Based Method' is used to refer to the latter procedure.

2.8.4.2.1 The Predictive Model Based Method

SOLAS Predictive Model Based Method uses linear regression (OLS) on each variable with missing values in turn, as described in Appendix C of the SOLAS user reference manual (Statistical Solutions, 1999). In the description which follows, the theoretical properties are based on the assumption that y has a Normal distribution.

Let y be a variable with missing values which are to be imputed, and let $x_1,\ldots,x_p$ be the p designated covariates of the variable y, where all values of $x_1,\ldots,x_p$ are observed for all cases. It is assumed that y depends on the covariates according to a linear regression model with Normally-distributed error term

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2)$$

This has $(p+1)$ coefficients, including β_0 for the intercept.

First, the regression coefficients $\beta_0, \beta_1, \cdots, \beta_p$, the variance σ^2 of the error term, and the covariance matrix $\mathbf{V}$ of the estimates of the βs are estimated by standard least-squares regression using as data only those cases for which y is observed.

This gives estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ of the βs, which have a multivariate Normal distribution with means $\beta_0, \beta_1, \ldots, \beta_p$ and covariance matrix estimated by **V**; and the estimate

$$\hat{\sigma}^2 = \frac{RSS}{n-p-1}$$

of σ^2 (where RSS is the sum of squares of the residuals from the regression). In making the imputation, the uncertainties due to the fact that these are estimates and not exact values is incorporated by making use of the multivariate Normal distribution of the βs, and the fact that RSS/σ^2 has a χ^2 distribution with $n-p-1$ degrees of freedom.

For the true variance σ^2 of the error term, the Bayesian posterior distribution can be derived from the above by rearranging it as follows:

$$\frac{RSS}{\sigma^2} \sim \chi^2_q \Longrightarrow \sigma^2 \sim \frac{RSS}{U} = \frac{n-p-1}{U} \times \frac{RSS}{n-p-1} = \frac{n-p-1}{U} \times \hat{\sigma}^2$$

where $U \sim \chi^2_q$, $q = n-p-1$.

Hence the Bayesian posterior for σ^2 can be sampled from by first sampling U from a χ^2 distribution with $q = n-p-1$ degrees of freedom, and then multiplying $\hat{\sigma}^2$ (estimated from the data) by $(n-p-1)/U$. Let this value be denoted by $\tilde{\sigma}^2$.

For the βs, the Bayesian posterior distribution with a uniform prior is multivariate Normal, with means equal to the $\hat{\beta}$s (estimated βs), and covariance matrix equal to

$$\mathbf{W} = \mathbf{V} \times \frac{\hat{\sigma}^2}{\tilde{\sigma}^2}$$

Hence, in making the imputation for a case of missing y, first U is sampled from a χ^2 distribution with $n-p-1$ degrees of freedom, then $\tilde{\sigma}^2$ is calculated, then **W** is obtained, and a sample of size 1 is drawn from the multivariate Normal distribution whose means are $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p$; the result of this sample is a set of values for $\beta_0, \beta_1, \cdots, \beta_p$; let these be denoted by $\tilde{\beta}_0, \tilde{\beta}_1, \cdots, \tilde{\beta}_p$.

Then a random value is drawn from a standard Normal distribution $N(0,1)$, and this is multiplied by $\tilde{\sigma}$ to give a value $\tilde{\epsilon}$ for the error term in the regression equation.

Finally, the values $x_1, \ldots, x_p$ of the covariates for the case where y is missing are used to generate an imputed value $\tilde{y}$ according to

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \cdots + \tilde{\beta}_p x_p + \tilde{\epsilon}$$

2.8.4.2.2 SOLAS Discriminant Method

SOLAS Discriminant Method imputes missing values of categorical variables by sampling from a Bayesian posterior distribution which assumes that the covariates have a multivariate Normal distribution where both the mean $\boldsymbol{\mu}_j$ and the covariance matrix $\boldsymbol{\Sigma}_j$ may depend on the value j of

the categorical variable. The following description is closely modelled on Appendix D of the SOLAS user reference manual (Statistical Solutions, 1999).

Let $1,\ldots,s$ be the categories of the categorical imputation variable y. By applying Bayes's theorem, the statistical model of discriminant imputation is given by the following equation:

$$P(y=j|\mathbf{x}) = \frac{\phi(\mathbf{x}|\mu_j;\Sigma_j)\pi_j}{\sum_{v=1}^{s}\phi(\mathbf{x}|\mu_v;\Sigma_v)\pi_j}, \quad j=1,\ldots,s.$$

In this equation $P(y=j|\mathbf{x})$ is the probability that the imputation variable y is equal to its jth category given the vector $\mathbf{x}$ of the observed values of the covariates of y, and $\phi(\mathbf{x}|\mu;\Sigma)$ is the density of the multivariate Normal distribution with mean μ and covariance matrix Σ; μ_j and Σ_j are the conditional mean and covariance matrix of the covariates of y given that y is equal to its jth category, and π_j is the prior probability that y is equal to its jth category.

The imputation scheme for discriminant mutiple imputation is given by:

(i) Let n_j be the number of observed values of y equal to the jth category of y, and let $a_j = \frac{1}{2} + n_j$ for $j=1,\ldots,s$.

(ii)

Draw $\theta_1^*,\ldots,\theta_s^*$ from the standard Gamma distribution with parameters given by $a_1,\ldots,a_s$.

(iii)

Let $\pi_j^* = \frac{\theta_j^*}{\sum_{v=1}^{s}\theta_v^*}$ for $j=1,\ldots s$.

(iv)

For $j=1,\ldots,s$, draw μ_j^* from the multivariate Normal distribution with mean $\hat{\mu}_j$ and covariance matrix S_j/n_j, where $\hat{\mu}_j$ and S_j/n_j are the sample mean and covariance matrix of the covariates of y calculated from the cases where y is equal to its jth category.

(v)

Let $p_{ij}^* = \frac{\phi(X_i^{\mathrm{T}}|\mu_j^*;S_j)\pi_j^*}{\sum_{v=1}^{s}\phi(X_i^{\mathrm{T}}|\mu_v^*;S_v)\pi_v^*}$, for $i=1,\ldots,n_{\mathrm{miss}}$ and $j=1,\ldots,s$, where $\phi(\mathbf{x};\mu,\Sigma)$ is the probability density function of the multivariate Normal distribution with mean μ and covariance matrix Σ, the index i denotes the ith missing value of y, k is the number of covariates used for the imputation variable y, and X_i^{T} is the row vector of covariates of y corresponding to the ith missing value of y.

(vi)

Let y_i^* be randomly drawn so that it is equal to j with probablity p_{ij}.

(vii)

Impute y_i^* for the ith missing data entry of y for $i = 1, \ldots, n_{\text{miss}}$.

2.8.5 Schafer's procedures

Joe Schafer (Schafer 1997) has produced software for multiple imputation which is freely downloadable from his website (**http://www.stat.psu.edu/~jls/misoftware.html**). These were originally written as libraries for S-plus versions 3.0, 4.0 and 4.5. In July 2001, S-plus 6 was released with a library 'missing' which provides the same resources, and Schafer's routines underlie this S-plus library. In their original form they have been slightly modified so as to be libraries for R (the Free Software equivalent to S-plus: see **http://www.stats.bris.ac.uk/R**).

Schafer's original three procedures CAT, NORM and MIX correspond to the three types of model Loglinear ('Loglin'), Gaussian ('Gauss') and conditional Gaussian model (Cgm) in the S-plus 'missing' library.

Each of these libraries is aimed at a particular type of data:

CAT is intended for use when all variables are categorical, i.e. are factors with discrete levels.

NORM is intended for use when all variables are continuous and may (possibly after a preliminary transformation) be assumed to have a multivariate Normal distribution.

MIX is for use when some variables are categorical and others are continuous. A multinomial distribution is assumed for the categorical data in which each combination of levels may have its own probability. The continuous variables are assumed to have a multivariate Normal distribution within each combination of levels of the categorical variables. Each combination of categories is associated with a multivariate Normal distribution. These distributions may have unrelated means, but have a common covariance matrix.

In each case the basic procedure is as follows. First, a preliminary sorting and summarizing of the data is done, to rearrange them into an order as near to monotone missingness (Section 2.7.1) as possible (for greater efficiency of computation) and to pre-compute the statistics that will be repeatedly used in later steps.

Next, the parameters of the probability distribution of the data (multinomial for categorical, multivariate Normal for the continuous) are estimated by maximum likelihood using the EM algorithm (Dempster, Laird and Rubin 1977), described in section 2.7.3. The imputations can then be made by filling in missing data case by case with values sampled from a distribution which depends on these parameter-values and is conditional on the values of the observed data for the case.

When the EM algorithm is used, the uncertainty in the parameter estimates is not readily available because it is not a direct by-product of the computational procedure. Schafer uses Data Augmentation (DA) (Schafer 1997, Chapter 4) for generating imputations. This procedure first samples parameter values from their Bayesian posterior distribution by MCMC (Markov

Chain Monte-Carlo), which only needs the likelihood function and a prior distribution to be available (as they are, directly, within Schafer's software routines); then, using the distribution with the sampled parameter values, imputed values are sampled as above. Multiple imputations are obtained by repeating the two stages of data augmentation. It is possible to draw multiple imputations by repeatedly sampling from the EM-estimated distribution without varying the parameter values.

For CAT and MIX it is possible to specify a model for the data, both at the stage of estimation by the EM algorithm and in the DA stage. In CAT, the structure of the log-linear model for the categorical data can be specified. The simplest model assumes independence of the variables (each cell probability is the product of the marginal probabilities). In the most complex (saturated) model, each combination of categories has a probability, and these probabilities are constrained only by the requirement that they add up to unity. In MIX, the same sort of specification as with CAT can be used for the categorical part of the data. In addition, a model can be specified for the dependence of the multivariate Normal means on the levels of the categorical variables. In NORM, the data are simply assumed to have a multivariate Normal distribution.

Markov Chain Monte Carlo

Markov Chain Monte Carlo (MCMC) is a straightforward and quite general algorithm for generating a sample from an arbitrary probability distribution. Distributions that arise in practice, especially Bayesian posterior distributions, may be far too complicated and intractable to be sampled from directly. The MCMC procedure only needs enough to be known about the probability density function $f(x)$ of the desired distribution to compute the ratio $f(u)/f(v)$ at any two points u and v. A random sequence (Markov chain) $x_1,\ldots,x_n,x_{n+1},\ldots$ is generated; given x_n, a possible next point y is randomly sampled from an 'easy' distribution with probability density $g(y|x_n)$, and then it is randomly decided, with a probability which depends on the ratio $\{\, g(x_n|y)\, f(y)\,\}/\{\, g(y|x_n)\, f(x_n)\,\}$, whether $x_{n+1} = y$ or $x_{n+1} = x_n$. As n increases, the distribution of x_n approaches the desired distribution $f(x)$, regardless of the starting point x_1. So, for n large enough, x_n can be taken as a value sampled from $f(x)$. To obtain a further value, the whole process is repeated.

The MCMC procedure is simple, and can be straightforwardly applied to practically any distribution. Its main difficulties are that n may need to be very large for the distribution of x_n to have converged sufficiently close to $f(x)$, and that testing whether n is large enough may be problematic. A general account covering both theory and applications can be found in Gilks et al. (1996).

2.9 Combining the results of multiply imputed datasets

The result of multiple imputation is m completed datasets, in which the missing values have been replaced by imputations, and the values of the imputations in each dataset may vary. To analyse the data, one uses the intended analysis (that would have been applied to the original data if they had been complete) to each completed each dataset in turn to give the required estimates, along with their respective standard errors. The m resulting estimates are combined to give a single result in the following way (Rubin, 1987).

Let the estimate of the scalar quantity Q of interest obtained from dataset j be $\hat{Q}_j$ $(j = 1,\ldots, m)$, with squared standard error U_j. The ***MI-estimate*** of Q is the average of the individual estimates:

$$\bar{Q} = \frac{1}{m}\sum_{1}^{m}\hat{Q}_j \tag{1}$$

The total variance consists of two parts: the within-imputation variance

$$\bar{U} = \frac{1}{m}\sum_{1}^{m}U_j \tag{2}$$

and the between-imputation variance

$$B = \frac{1}{m-1}\sum_{1}^{m}(\hat{Q}_j - \bar{Q})^2 \tag{3}$$

giving total variance

$$T = \bar{U} + \left(1 + \frac{1}{m}\right)B \tag{4}$$

$\bar{U}$ is the estimate of what the sampling variance of the estimate would have been had the data been complete. B is the additional variance due to the uncertainty about the missing values. B/m is the inflation of the 'between' variance, needed because only m imputations have been used. T is the MI-estimate of the sampling variance that takes account of the uncertainty due to missing data.

The following two diagnostic measures indicate how strongly the quantity estimated is influenced by missing data, or non-response (Rubin, 1987), by comparing the additional variance due to non-response with, respectively, the sampling variance or the total variance of the estimate.

First, the relative increase in variance due to non-response compares the variance between the imputations (B) with the estimated sampling variance ($\bar{U}$), and it is defined as

$$r = \frac{T - \bar{U}}{\bar{U}} = \frac{(1 + 1/m)B}{\bar{U}} \tag{5}$$

Second, the fraction of missing information (γ) about the estimate of Q due to non-response essentially compares the variance between the imputations (B) with the total variance (T) (Rubin, 1987, p. 21). This is estimated (Schafer, 1997, p. 110) as

$$\gamma = \frac{r + 2/(\mathrm{df} + 3)}{r + 1} \tag{6}$$

where

$$\mathrm{df} = (m - 1)\left(1 + \frac{m\bar{U}}{(m + 1)B}\right)^2 \tag{7}$$

The degrees of freedom (df) are large for large numbers (m) of imputations, or when the between-imputation variance (B) is small relative to the sampling variance ($\bar{U}$), in which case γ is approximately the ratio of the 'between' to the 'total' variance:

$$\gamma \approx \frac{r}{r + 1} = \frac{(1 + 1/m)B}{(1 + 1/m)B + \bar{U}} = \frac{(1 + 1/m)B}{T}$$

The relative efficiency (RE) is the approximate efficiency of using an m-imputation estimator instead of an infinite number for the fully-efficient imputation, given by

$$RE = \left(1 + \frac{\gamma}{m}\right)^{-1} \tag{8}$$

The relative efficiency depends on γ, but unless γ is very high there is little advantage in increasing m beyond a small number. For example, even if 50 per cent of the information is missing, 5 imputations give a relative efficiency of 91%, while doubling m to 10 only increases the RE to 95%. A high relative efficiency can be used as a post-hoc justification for the choice of the number m of imputations.

Chapter 3

Measuring Alcohol Consumption in the MRC National Survey of Health and Development

3.1 Introduction

Adverse health consequences are associated with excessive alcohol consumption (Section 1.2.1). Population averages do not inform us about the number of people who drink excessively (Section 1.2.2). In order to be able to investigate the relationship of alcohol consumption with health consequences we need to know how much individuals drink (Section 1.2.3).

Individual levels of alcohol consumption in the general population are derived from surveys in which a sample of subjects from the population is asked what they drink. The estimates derived from the data collected from these general population surveys invariably fall short of the estimates of national consumption based on excise duty. This is termed the problem of 'coverage' and generally survey estimates represent only about 40–60% of annual sales (Pernanen, 1974). In other words, self-reported alcohol consumption in general population surveys appears to dramatically underestimate consumption. Two factors are usually cited to account for this underestimation:

(i) under-reporting of alcohol consumption by respondents, and

(ii) low response rates in general population surveys of drinking.

Neither of these would present a problem if the resulting estimates were not differentially biased. Under-reporting could be adjusted for if all responders under-report to a similar degree by inflating the reported consumption by a multiplicative constant to make the total agree with that given by the sales figures. However, it is commonly believed that heavy drinkers tend to underestimate their alcohol consumption more than light drinkers because they are more likely to forget how much they have drunk, or because they are more likely to deliberately under-report their drinking from self-consciousness. In other words, self-reported consumption is believed to give estimates of the actual consumption level which have different biases for heavy and for light drinkers. Research evidence is far from clear, as Midanik's (1988) review of the literature shows. Some research supports this conjecture (for example, Poikolainen, 1985), while some provides evidence to the contrary (for example, Lemmens et al., 1988).

Although low response rates result in inefficiency, the estimates themselves are not biased if non-response is not related to drinking — that is, if those who do not respond drink at similar

levels to those who respond to the survey. However, it is believed that heavy drinkers are more likely to refuse to take part in surveys. If this is the case, alcohol consumption is systematically different for non-responders and responders and hence estimates based on the sample of responders are not representative of those in the general population. The evidence for non-response bias is often indirect because the alcohol consumption of non-responders is not available unless they can be contacted and the information collected in a follow-up. Indirect evidence is provided by observed characteristics of the non-responders that are related to alcohol consumption in responders. In many surveys, subpopulations with a higher proportion of heavier drinkers tend to show higher non-response rates (Pernanen, 1974). Direct evidence of heavy drinking amongst non-responders is seldom presented, however, because of the difficulty and expense of following up non-responders. The reason for non-response may be that subjects could not be contacted in the first place, rather than unwillingness to take part. Difficulty in contacting people has been found to be related to higher than average alcohol consumption (Wilson, 1981). In a Canadian survey, larger purchases of alcohol consumption were reported by respondents requiring several house calls than by respondents who were at home at the first call (De Lint, 1981). Where direct evidence has been collected from follow-up of non-responders the results have not always supported the conjecture that non-responders do drink more. Higher abstention rates have been found amongst non-responders than responders to alcohol surveys (Garretson,1983; Mulford and Miller, 1959). Knibbe (*see* Lemmens et al., 1988) reported higher rates of abstention in non-responders than responders in a Dutch alcohol survey, but also more frequent heavy drinking. Moreover, where non-responders are followed up, direct evidence about their alcohol consumption does not always agree with the indirect evidence based on their characteristics. For example, in a British survey on alcohol consumption (Crawford, 1986) female non-participants were more likely to be employed and to be non-manual workers, both factors related to higher alcohol consumption, but follow-up of these women did not support the inference that their alcohol consumption was high.

The MRC National Survey of Health and Development (NSHD) collected information about alcohol consumption when the survey members were 43 years of age (Section 2.3). The two sources for information about alcohol consumption are the weekly recall of the total number of alcoholic beverages and the seven-day diet diary. The seven-day diary required much more commitment on the part of the respondent since it involved recording all the food and drink consumed during each of the seven days of the week. As a consequence, many subjects did not complete the diary for the entire week. The survey also collected information on problems with drinking using the CAGE questionnaire.

In this chapter it is shown that the diary is a more valid source of information about alcohol consumption than the recalled weekly total. Section 3.2 examines these sources of information and the extent and nature of non-response to them. Section 3.3 reports the estimates of the prevalence of excessive alcohol consumption using recall and diary. Section 3.4 compares the validity of the diary with that of recall for measuring alcohol consumption. First it is argued

that there is a greater extent of 'under-reporting' in the recall measure compared to the diary, and then that the extent of the under-reporting in the recall compared to the diary is biased by the respondents' attitudes to their drinking as reported in answers to the CAGE questions.

3.2 The response patterns to recall, diary and CAGE

3.2.1 Summary measures of alcohol consumption and drink problems

In the NSHD interview there were two instruments used to collect information on alcohol consumption: 'weekly recall' (Section 2.3.2) and 'seven-day diary' (Section 2.3.3); and the CAGE questionnaire (Section 2.3.1) collected information on problems with drinking. Each of these instruments consists of several items. In practice, the responses to the items within each instrument are combined to give an overall summary measure of alcohol consumption (for recall and diary) or of drink problems (CAGE). The weekly recall is used to give the total alcohol consumed (in Units) in the week prior to the interview, by adding up the number of drinks given in response to each item: spirits, wine and beer. We can summarise the diary information in many different ways. In this chapter, the total alcohol consumed during the diary week is used for comparison with the weekly recall total. The diary week total is calculated by adding up the alcohol consumed on each of the seven diary days (converted to Units of alcohol as described in section 2.3.3). To get an overall CAGE score we add the number of affirmative answers to each of the four CAGE questions. The CAGE questionnaire was developed as a screening device for alcoholism in clinical settings (Ewing, 1984). More recently it has been increasingly used in general population surveys in Britain, in which a score of two or more is used to indicate that the subject has a drink problem (Hedges, 1996; Hope et al., 1998). We adopt the same criterion in the NSHD (Richards et al., 1997; Ely et al., 1999). A CAGE score of two or more on questions about problems in the previous year (CAGE LAST YEAR, Section 2.3.1) is used here to indicate current problems with drinking.

3.2.2 Total and partial non-response

The summaries described in Section 3.2.1 require complete responses to all the items within the instrument, or complete data for that instrument. Some respondents may not complete any of the items (the record is empty), when, with respect to the instrument, non-response may be said to be 'total'. Others may complete some, though not all, of the items, when non-response can be said to be 'partial', so any respondent who did not complete all the items could not be assigned a summary value. For example, with respect to the weekly recall instrument, non-response is total for respondents who recorded the quantities for neither beer, nor wine nor spirits; and non-response is partial for respondents who failed to record the amount of beer consumed in the last week, whilst recording the amount of wine and spirits. Non-response to the diary instrument is total if no diary days are completed, and partial if less than seven days are completed. To give the extent of non-response in the diary we first discuss the details of the diary structure.

3.2.3 The structure of the diet diary data

From the diet diary, we derive seven items corresponding to the alcohol consumption recorded for each of the days of the diary week. It is important to appreciate that if the diet diary is completed for a particular day, then so is the alcohol consumption. This is because the subjects were asked to record in the diary all food, including alcoholic and non-alcoholic beverages. Although the reminder section prompted the subject to record any beer, wine, sherry or spirits not previously recorded, this was to be recorded only if it was applicable. Therefore anyone who did not record any alcoholic beverages was assumed to have consumed none, so from any diary days returned we can deduce the quantity of alcohol consumption recorded on that day. The way this is done is described in Section 2.3.3.

Not all subjects completed all the seven diary days. In Table 3.1, the second line shows the number of subjects completing a given number of the days. For example 892 subjects completed only two days.

Table 3.1: Missing data in the seven-day diary records

Days completed	**7**	**6**	**5**	**4**	**3**	**2**	**1**	**0**
Number of subjects	2089	77	96	8	13	892	11	76
Number of days (*n*)	**7**	**6**	**5**	**4**	**3**	**2**	**1**	**0**
Number of subjects completing *n* days on schedule	2002	71	81	10	17	970	14	97
Number of subjects completing at least *n* days on schedule	2002	2073	2154	2164	2181	3151	3165	3262
% of subjects completing at least *n* days on schedule	61.4	63.6	66.1	66.4	66.9	96.6	97.0	100.0

As discussed in Section 2.3.3, the diaries started two days before the nurse interviewed the respondent, and subjects were asked to fill in a diary sheet for each of the following five days. For each such day a diary sheet filled in according to this instruction is said to be completed 'on schedule'. Some subjects ($n = 108$) omitted one or more days and completed the diary on later days, thus not completing the days according to the instructions, that is, not on schedule. Such entries may be biased because of the influences that trigger the lapse and re-continuation of the diary: for example, some subjects went away on holiday, and continued their diary on their return. (The subjects were sent no reminders.) The diary records for these subjects were not totally discarded since they had (except for $21 = 97 - 76$) partially completed the diary as instructed. Only the 541 days (2.4%) recorded on days later than scheduled were discarded,

giving the number of days completed on schedule, as shown in the second part of Table 3.1.

The diary came at the end of a two and a half hour interview and 97 interviews were terminated without starting the diary. Only 2002 (61.4%) subjects completed their diaries on schedule. However, the proportion of subjects with partial response was high. Some subjects (970) completed the first two days with the nurse, but failed to complete any subsequent days on their own or failed to mail back their diary. Altogether 3151 or 96.6% recorded at least two diary days. The patterns of responses on schedule are shown graphically in Figure 3.1. They have a monotone missing pattern (defined in Section 2.7.1).

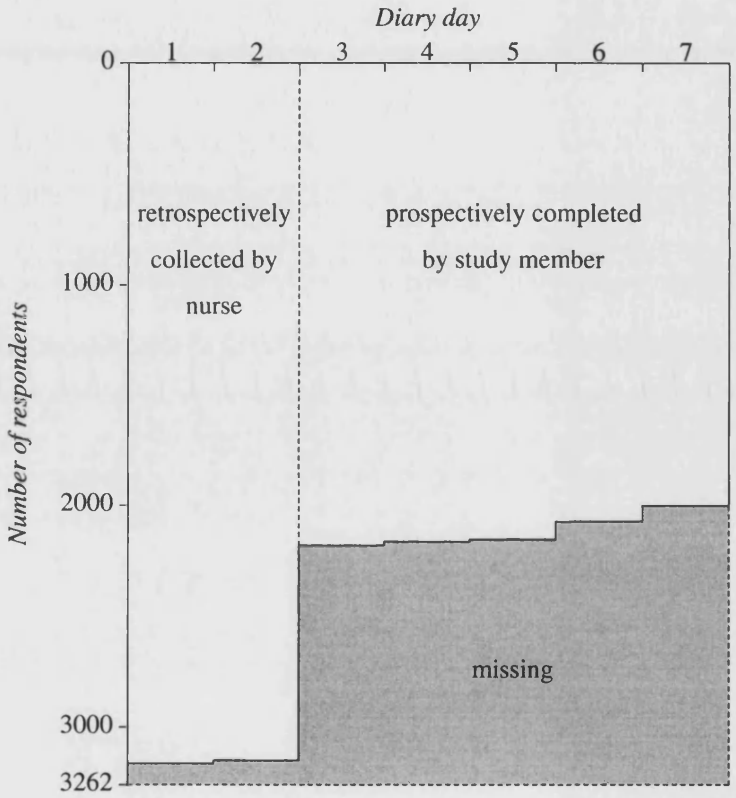


Figure 3.1: Missing data in the seven-day diet diary

3.2.4 The extent of non-response

The numbers (%) of subjects who completed the recall, diary, and CAGE questionnaire, and the extent of partial and total non-response to these instruments (defined in Section 3.2.2), are given in Table 3.2 below.

The extent of total non-response was small, and similar for recall (2.7%), diary (3.0%) and CAGE LAST YEAR (2.1%). Total non-response to CAGE EVER was greater: six percent of the respondents answered none of the questions relating to life-time experience of problems with drinking. The difference between CAGE EVER and CAGE LAST YEAR might be explained by the layout of the questions. The questions relating to the last year were on the

Table 3.2: The extent of partial and total non-response to recall, diary and CAGE

	RECALL		DIARY		CAGE LAST YEAR		CAGE EVER	
	n	%	*n*	%	*n*	%	*n*	%
Complete data	2456	75.3	2002	61.4	3169	97.1	2942	90.2
Partial non-response	719	22.0	1163	35.7	25	0.8	125	3.8
Total non-response	87	2.7	97	3.0	68	2.1	195	6.0
Total	3262		3262		3262		3262	

right hand side of the sheet of paper and could have been more visible to the respondent (see Appendix 1).

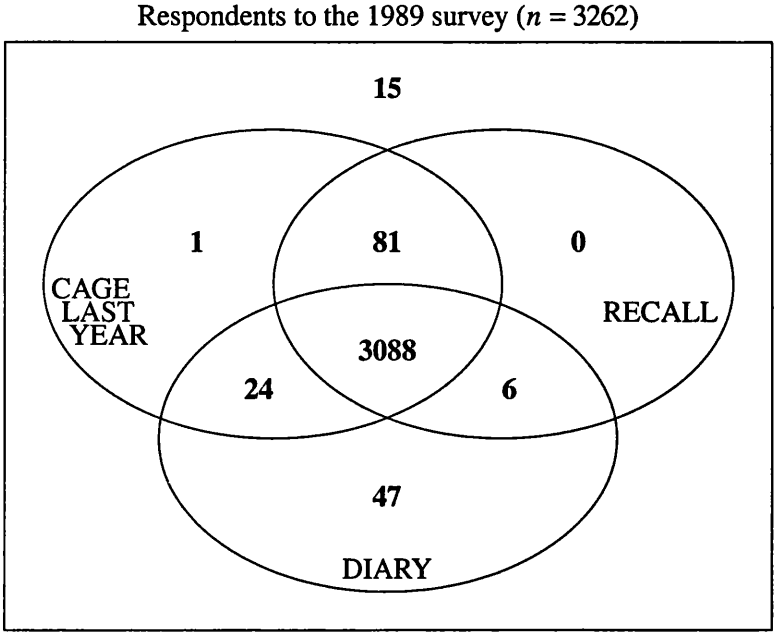
For the instruments measuring alcohol consumption (recall and diary), the extent of partial non-response was much greater than that of total non-response, and much greater than partial non-response to CAGE. Besides those who did not answer any of the items in the instruments, an additional 22.0% omitted at least one of the types of drink in the recall and 35.7% (1163/3262) failed to record at least one of the diet diary days on schedule, whilst only 0.8% omitted at least one of the CAGE LAST YEAR questions.

Although the extent of partial response to the instruments used to measure alcohol consumption presents a more serious problem than total non-response, information about the respondents could be obtained from the items that had been completed. For those with total non-response to an instrument, information could be provided by the alternative instruments. The overlap in (at least) partial response to combinations of items in the recall, diary and CAGE LAST YEAR is given in Figure 3.2 below. Subjects who failed to respond totally to one of the three instruments generally responded to another, so that some information about drinking was available for all but 15 cases (less than 0.5%) of the respondents at age 43 (see Figure 3.2). In only 16 cases was there was no information about quantity of alcohol consumed. Most respondents (3094 or 94.3%) provided some information for both the recall and the diet diary items (see Figure 3.2).

3.2.5 Influences on non-response

In order to assess the likely impact of missing values on the estimates produced by a particular survey we first consider evidence relating to the specific circumstances of the survey. In this survey all the items referring to alcohol came at the end of a long interview and in some cases the nurse was unable to complete the survey because of pressure of time. The diet diary instructions and recording of the first two days of the diary constituted the final part of the long structured interview. (The mean length of the interview was 2 hours 12 minutes and 5% of interviews were more than two and a half hours long.) The recall and CAGE items were asked at the end of a self-completed questionnaire which participants were asked to fill in during the interview. The diary was left with the survey member who then had to complete the remaining five days prospectively and forward it in an envelope provided to the study team, without further reminders. While this survey methodology ensured a high level of response to the first two

Figure 3.2: Numbers of subjects interviewed in 1989 responding (at least partially) to combinations of items in the recall, diary, and CAGE LAST YEAR



diary days (96.6%), it also explains the high level of partial non-response, namely that 33% (1081/3262) of the respondents failed to return their diary (see Table 3.1). There may be many reasons that are unrelated to the drinking behaviour for the failure to return the diary. Since the diet diary was not designed to collect information specifically about alcohol consumption and non-response to this item consisted of a failure to complete sheets on general dietary consumption, diary non-response is not necessarily related to the alcohol consumption of the respondent. How non-response to the diet diary relates to other information collected from study members is examined in Chapter 5.

While partial missing responses to the diet diary may be expected because of the commitment that such an instrument required of the respondents, the high level of partial non-response to the concise recall questions calls for some other explanation. A plausible explanation of non-response to the specific questions on alcohol contained in the recall and CAGE items is that respondents who did not drink disregarded the questions altogether because they found them irrelevant. In this survey abstainers (people who do not drink alcohol at all) were not identified specifically by the questionnaire and questions about recalled consumption and problems were directed at all respondents (see Appendix 1). The two previous sections of the self-completion part of the questionnaire had stated that they were 'for everybody' whereas the final section which contained the drink questions made no reference to whom they applied, so abstainers may have assumed that these questions did not apply to them.

Women are generally more conscientious in answering survey questions than men. For example, women were more likely than men to complete their diary (62.9% of women vs. 59.8% of men). Yet women were less likely than men to respond to CAGE EVER (7.5% of women and 4.5% of men left all the CAGE items blank) or recall questions (27.8% of women and 21.7% of men did not complete their recall). Women are also more likely to be abstainers than men. All the 128 respondents who left the CAGE EVER questions blank, but who answered the 'last year' questions, denied having problems in the last year. Of these, all who answered questions about recalled consumption in the previous week reported drinking no alcohol. For some respondents with total non-response to the CAGE, the research nurse interviewers provided written comments on the questionnaire sheet that the respondent had not completed the CAGE questions because they did not drink.

There is strong evidence that subjects with partial non-response to the recall did not enter a zero when they had not drunk that particular type of alcoholic beverage, perhaps because they never drank it. Inspection of the questionnaire revealed that some subjects who were coded as non-responders to a particular item had put a line through the box provided on the questionnaire. Although 25% (806) of the respondents left at least one of the recall questions blank, only 0.6% (19) used both blank and a zero. In addition, the distribution of item non-response by gender and social class was similar to that of the zero entries. For example, women were more likely than men to leave their beer consumption blank, particularly women from higher social classes, and these are exactly the groups who are unlikely to drink beer. The reverse was true for wine, which is drunk more by women than by men, and more by those in higher than lower social classes. In the light of this evidence, partial non-response to the recall or CAGE was interpreted as adding nothing to the total amount drunk in the recall. As a result of adopting this strategy for dealing with partial non-response, 3175 (97.3%) subjects could be assigned a total alcohol consumption using the recall instrument. Using the diary instrument only the 2002 respondents who completed their diary could be assigned a total alcohol consumption.

3.3 Estimates of the prevalence of excessive alcohol consumption using recall and diary

The mean of the total alcohol consumption reported in a weekly period based on the recall is 9.2 Units ($n = 3175$) while the diary gives a higher mean of 12.8 Units ($n = 2002$). However, comparing means is not very informative because the distribution of alcohol consumption is very positively skewed (see Section 2.4), so the mean is affected by the few respondents who drink very heavily. The level of alcohol consumption is generally classified according to categories, rather than being treated as a continuous measure. The categories used in this dissertation refer to those in common use in health promotion literature at the time the respondents were interviewed (Health Education Council, 1984, 1985) derived from contemporary research on the harmful effects of alcohol consumption (Royal College of General Practitioners, 1986; Royal College of Psychiatrists, 1986). Respondents are classified as

'sensible', 'immoderate' or 'heavy' drinkers according to the accepted gender-specific criteria (Royal College of Psychiatrists, 1986), as shown in Table 3.3.

Table 3.3: Classification of drinkers according to reported consumption levels

	Level of weekly consumption in Units	
Description	**Men**	**Women**
Sensible	up to 21	up to 14
Immoderate	over 21, up to 50	over 14, up to 35
Heavy	over 50	over 35

As this is a classification of drinkers, abstainers are generally excluded from the category of 'sensible' drinkers. However, as we have seen, abstainers cannot be identified in the NSHD survey (Section 3.2.5). Nevertheless it is of interest to examine those reporting zero alcohol consumption as a separate category in this instance; this level is described as 'None' in Table 3.4.

Table 3.4: Recall and diary instruments — percentage of respondents reporting weekly alcohol consumption in categories of alcohol consumption

	Recall				Diary			
	Men		Women		Men		Women	
Level	*n*	*%*	*n*	*%*	*n*	*%*	*n*	*%*
None	227	14.2	465	29.5	145	14.8	288	28.1
Sensible	1048	65.5	1013	64.4	508	51.9	576	56.3
Immoderate	263	16.4	88	5.6	241	24.6	149	14.6
Heavy	63	3.9	8	0.5	84	8.6	11	1.1
ALL	1601		1574		978		1024	

Table 3.4 gives the percentages of respondents reporting levels of alcohol consumption according to each of the two instruments by those subjects who completed the instrument. Whereas there is close agreement between the instruments in the proportions of men and women who take no alcohol, the diary gives a smaller proportion of sensible drinkers, and greater proportions of immoderate and heavy drinkers, than the recall. The focus of interest from the health perspective is on excessive levels of alcohol consumption. This dissertation is concerned with the estimation of the proportions of people drinking above recommended levels: those drinking above sensible levels, or ***excessively*** (men drinking more than 21 Units per week and women more than 14 Units per week), and those drinking above moderate levels, or ***heavily***

(men drinking more than 50 Units and women more than 35 Units per week). The two instruments give very different estimates for these proportions (Table 3.5).

Table 3.5: Comparison of recall and diary instruments: estimates of the percentage of respondents reporting weekly alcohol consumption above weekly limits (Units)

	Recall					Diary				
	n	%	se	95% CI		*n*	%	se	95% CI	
Women	1574					1024				
>14 U	96	6.1	0.60	(4.92	7.28)	160	15.6	1.13	(13.40	17.85)
>35 U	8	0.5	0.18	(0.16	0.86)	11	1.1	0.32	(0.44	1.71)
Men	1601					978				
>21 U	326	20.4	1.01	(18.39	22.33)	325	33.2	1.51	(30.28	36.18)
>50 U	63	3.9	0.49	(2.98	4.89)	84	8.6	0.90	(6.83	10.35)

Let us assume, in this instance, that the people who provide complete data are representative of all those interviewed in 1989, and inferences are restricted to the population well represented by the 3262 subjects. Under these assumptions, estimating, for example, the proportion of women drinking excessively (more than 14 Units per week) using the diary instrument yields 15.6% with a 95% confidence interval of 13.40 to 17.85; using the recall instrument yields an estimate of only 6.1% (4.92 to 7.28). All the estimates are substantially, and significantly, higher based on the diary than on the recall, except for women who drank heavily (more than 35 Units), of whom there are very few for either the recall (8 subjects) or the diary (11 subjects).

3.4 Validity of the diary instrument for measuring alcohol consumption

3.4.1 Under-reporting in the recall relative to the diary

Comparison of the results from the two instruments in Table 3.5 provides evidence of under-reporting alcohol consumption in the recall measure relative to the diary. It is not obvious how one should measure the extent of this 'under-reporting'. One indicator is the difference between the total weekly consumption given by the two measures. Using this indicator, under-reporting increases with consumption level because people who drink a lot tend to have larger differences. For example, an extra 10 Units in the diary total is more likely, and less 'significant', for someone who drinks 100 Units in a week compared with someone who only drinks 20 Units. Although the additive difference in reported consumption is the same in each example, the first case represents a proportional under-reporting of only 10%, compared with 50% in the latter case. If proportional differences are used the opposite is true: proportional differences in the two measures decrease with consumption: people who drink little tend to have large proportional differences. For example, someone who reports drinking one pint of beer in the recall but eight pints in the diary has multiplied their drinking by a factor of eight, the same as someone

who reports drinking 10 pints of beer in recall and 80 pints in the diary. Measuring under-reporting either by additive or by proportional differences is not satisfactory. Motivated by public health information on the implications of levels of drinking, we base comparisons on the classification of drinkers as 'sensible', 'immoderate' or 'heavy' as shown in Table 3.3. Respondents are classified according to their reported consumption in both the recall and diary measures. Agreement between the two instruments is measured by the agreement in these classifications. The proportion of respondents who are in a lower category in their recall than in their diary is used as a measure of under-reporting in the recall relative to the diary.

We cannot expect that a subject would report the same alcohol consumption in the recalled week and the diary week because the two measures do not relate to the same period. The comparison of the two measures is affected by the variation in consumption from week to week, although this will be reduced since the periods covered by the recall and the diary overlap by two days. If the difference between the recall and diary totals is entirely due to the variation in levels of drinking from week to week, the proportions of respondents with positive and negative differences between diary and recall consumption would differ only due to chance. There would be an approximately equal proportion of respondents with positive and negative values for the difference in diary and recall consumption. But of the 1962 respondents who recorded both their weekly recall and completed the seven-day diary, only 24.1% declared more in the recall than in the diary, whilst 59.9% declared more in the diary than in the recall. The difference in these proportions, 35.8% with a 95% confidence interval (32.3,39.3), is significantly greater than zero. Exact agreement between the two measures arises only for those who declared no consumption in both instruments. 15.7% reported drinking no alcohol on both instruments: 9.8% of the 964 men and 21.4% of the 998 women. These subjects are likely to drink only infrequently or not at all. Because they cannot be assumed to be abstainers they are included in the 'sensible' drinkers in the following comparisons (Tables 3.6 through 3.8).

The instruments are now compared according to their classification of respondents as sensible, moderate or heavy drinkers (as defined in Table 3.3). The results for the men and women who completed both the diary and the recall are given in Table 3.6. Only 2.3% (22/964) of all the men and 1.3% (13/998) of all the women were classified in a lower category by their diary than by their recall; in contrast, 21.5% (207/964) of men and 11.4% (114/998) of women were classified in a higher category by their diary than by their recall. Most of the latter were classified as sensible drinkers by their recall, but as immoderate drinkers by their diary total (19.1% of men and 11.2% of women who recalled drinking at a 'sensible' level). In addition, almost thirty percent (45/151) of men classified as immoderate drinkers by their recall were classified as heavy drinkers by their diary declarations.

Table 3.6: Classification of respondents' drinking level according to recall and diary totals

Men

		Recall						
		Sensible		Immoderate		Heavy		All
		%	*n*	%	*n*	%	*n*	%
Diary	**Sensible**	79.3	622	13.2	20	0.0	0	66.6
	Immoderate	19.1	150	57.0	86	6.9	2	24.7
	Heavy	1.5	12	29.8	45	93.1	27	8.7
	All	81.3	784	15.7	151	3.0	29	$n = 964$

Women

		Recall						
		Sensible		Immoderate		Heavy		All
		%	*n*	%	*n*	%	*n*	%
Diary	**Sensible**	88.6	829	19.0	11	0.0	0	84.2
	Immoderate	11.2	105	69.0	40	50.0	2	14.7
	Heavy	0.2	2	12.1	7	50.0	2	1.1
	All	93.8	936	5.8	58	0.4	4	$n = 998$

In conclusion, the differences between reported consumption in the diary and in the recall could not be accounted for by variations in drinking from week to week. The reported consumption in the diary tends to exceed that reported in the recall, and the differences have a substantial effect on estimates of the prevalence of excessive drinking.

3.4.2 Attitudes to drinking and under-estimation of alcohol consumption

We now explore how the extent of under-reporting in the recall relative to the diary is related to responses to the CAGE instrument (Section 3.2.1). Comparison of underestimation for those who responded affirmatively compared to those who did not was used to indicate the sensitivity of the instruments for measuring alcohol consumption to attitudes to drinking.

Table 3.7 gives the classification of the respondents' level of alcohol consumption by the recall and the diary for those with and without drink problems. More men report drink problems than women: 10.8% (104/962) of men and 4.8% (48/998) of women. Both men and women with drink problems (CAGE scores of 2, 3, or 4) drank more heavily than those without (CAGE scores of 0 or 1), whether their alcohol consumption was measured using the recall or the diary.

Table 3.7: Differences between recall amd diary in the classification of respondents' level of alcohol consumption for those without and with drink problems

Men

Without drink problems (CAGE score 0 or 1)

		Recall							
		Sensible		Immoderate		Heavy		All	
		%	*n*	%	*n*	%	*n*	%	*n*
Diary	Sensible	81	597	13	14	0	0	71	611
	Immoderate	18	129	59	61	6	1	22	191
	Heavy	1	11	28	29	94	16	7	56
	all: row%; *n*	86	737	12	104	2	17	100	858

Men

With drink problems (CAGE score 2, 3 or 4)

		Recall							
		Sensible		Immoderate		Heavy		All	
		%	*n*	%	*n*	%	*n*	%	*n*
Diary	Sensible	51	23	13	6	0	0	28	29
	Immoderate	47	21	53	25	8	1	45	47
	Heavy	2	1	34	16	92	11	27	28
	all : row%; *n*	43	45	45	47	12	12	100	104

Women

Without drink problems (CAGE score 0 or 1)

		Recall							
		Sensible		Immoderate		Heavy		All	
		%	*n*	%	*n*	%	*n*	%	*n*
Diary	Sensible	90	813	25	11	0	0	87	824
	Immoderate	10	91	73	32	100	2	13	125
	Heavy	0	0	2	1	0	0	0	1
	all: row%; *n*	95	904	5	44	0	2	100	950

Women

With drink problems (CAGE score 2, 3 or 4)

		Recall							
		Sensible		Immoderate		Heavy		All	
		%	*n*	%	*n*	%	*n*	%	*n*
Diary	Sensible	50	16	0	0	0	0	33	16
	Immoderate	44	14	57	8	0	0	46	22
	Heavy	6	2	43	6	100	2	21	10
	all: row%; *n*	67	32	29	14	4	2	100	48

In this table, the cells that indicate where recall underestimates alcohol consumption relative to the diary are shown shaded in grey. The sum of the numbers in these shaded cells gives the number of those who underestimate their alcohol consumption in the recall relative to the diary, and the proportions who do so are given in Table 3.9. Respondents with drink problems are more likely to underestimate their drinking in the recall relative to the diary than those who do not. For the men with drink problems the estimated proportion is 36.5% (with a 95% confidence interval of 27.3 to 45.8) for those with drink problems, compared with 19.7% (17.0,22.4) for those who do not have drink problems. The differences for women with and without drink problems are even more extreme (45.8% compared with 9.7%).

The CAGE questionnaire consists of the four questions '*C*ut down', '*A*nnoy', '*G*uilty' and '*E*yeopener' (Section 2.3). Apart from the question concerning the use of an '*E*ye-opener' the CAGE questions reflect psychological attitudes to drinking. The question most frequently answered in the affirmative was '*C*ut down': 22.5% of the men and 9.7% of the women responding to this question thought they ought to cut down on their drinking, followed by '*G*uilty' (9.3% of men and 4.9% of women felt guilty about their drinking), '*A*nnoy' (8.2% of men and 3.1% of women had been annoyed by people criticising their drinking) and least often '*E*ye-opener' (1.4% of men and 0.6% of women had needed a drink first thing in the morning). Women with a CAGE score of 2 were somewhat more likely to have felt *G*uilty (77%) than men (70%).

The numbers of respondents underestimating their alcohol consumption in the recall relative to the diary is given for each CAGE question individually in Table 3.8, and the proportions summarised in Table 3.9 in the same way as for the overall CAGE score which identifies those with drink problems.

Respondents who had felt they ought to *C*ut down on their drinking (in the past year) were more likely to underestimate their drinking using recall than those who had not felt this way. This was true of men and women alike. Respondents who answered affirmatively to *A*nnoy and *G*uilty were also more likely to underestimate their drinking in the recall relative to the diary. For women, the largest proportion underestimating their drinking were those who felt guilty about their drinking. Almost half of the women (46.9%; 95% CI 33.0% to 60.9%) who felt guilty about their drinking underestimated it compared with only 9.5% (95% CI 7.6% to 11.4%) who did not feel this way. Interestingly, responses to the *E*ye-opener question, the one question which is related to physical symptoms rather than psychological attitude, did not seem to be associated with changes in categories of drinker. However, the numbers of men ($n = 13$) and women ($n = 6$) who responded affirmatively to this question were too small to support any conclusions.

Overall, men were more likely to underestimate their drinking than women (Table 3.7), but the association with the CAGE was not as dramatic for men as it was for women (Table 3.9). Although the recall measure underestimates consumption, the recorded quantities could be

adjusted for if the extent of under-reporting was similar for all respondents. However, the extent of under-reporting was associated with respondents' attitudes to drinking. Those who reported having problems with drinking in the CAGE questions were more likely to under-report their consumption in recall.

[Text continued following Tables 3.8 and 3.9 below]

Table 3.8: Differences in the classification of respondents' drinking level according to recall and diary total by individual CAGE questions

Counts	Men						Women					
	Responded 'no' to 'Cut down'			Responded 'yes' to 'Cut down'			Responded 'no' to 'Cut down'			Responded 'yes' to 'Cut down'		
		Recall			Recall			Recall			Recall	
Diary	Sensible	Immoderate	Heavy	Sensible	Immoderate	Heavy	Sensible	Immoderate	Heavy	Sensible	Immoderate	Heavy
Sensible	565	10	0	53	10	0	800	8	0	28	3	0
Immoderate	100	42	0	50	43	2	75	20	2	29	19	0
Heavy	9	12	8	3	33	19	0	0	0	2	7	2
	Responded 'no' to 'Annoy'			Responded 'yes' to 'Annoy'			Responded 'no' to 'Annoy'			Responded 'yes' to 'Annoy'		
		Recall			Recall			Recall			Recall	
Diary	Sensible	Immoderate	Heavy	Sensible	Immoderate	Heavy	Sensible	Immoderate	Heavy	Sensible	Immoderate	Heavy
Sensible	601	19	0	18	1	0	816	11	0	12	0	0
Immoderate	132	69	1	17	17	1	98	39	2	7	1	0
Heavy	11	36	20	1	9	7	1	6	1	1	1	1
	Responded 'no' to 'Guilty'			Responded 'yes' to 'Guilty'			Responded 'no' to 'Guilty'			Responded 'yes' to 'Guilty'		
		Recall			Recall			Recall			Recall	
Diary	Sensible	Immoderate	Heavy	Sensible	Immoderate	Heavy	Sensible	Immoderate	Heavy	Sensible	Immoderate	Heavy
Sensible	595	15	0	25	5	0	812	11	0	16	0	0
Immoderate	128	68	2	19	18	0	88	32	2	16	8	0
Heavy	11	35	18	1	10	8	1	1	0	1	6	2
	Responded 'no' to 'Eyeopener'			Responded 'yes' to 'Eyeopener'			Responded 'no' to 'Eyeopener'			Responded 'yes' to 'Eyeopener'		
		Recall			Recall			Recall			Recall	
Diary	Sensible	Immoderate	Heavy	Sensible	Immoderate	Heavy	Sensible	Immoderate	Heavy	Sensible	Immoderate	Heavy
Sensible	615	20	0	4	0	0	825	11	0	4	0	0
Immoderate	149	83	2	1	3	0	104	40	2	1	0	0
Heavy	12	45	22	0	0	5	2	6	2	0	1	0

Table 3.9: Proportions underestimating consumption in the recall relative to the diary by gender and CAGE questionnaire responses

	Men			Women		
	n	%	95% CI	*n*	%	95% CI
Drink problems — No	858	19.7	(17.0 22.4)	950	9.7	(7.8 11.6)
Drink problems — Yes	104	36.5	(27.3 45.8)	48	45.8	(31.7 59.9)
***C**ut down* — No	746	16.2	(13.6 18.9)	905	8.3	(6.5 10.1)
***C**ut down* — Yes	213	40.4	(33.8 47.0)	90	42.2	(32.0 52.4)
***A**nnoy* — No	889	20.1	(17.5 22.8)	974	10.8	(8.8 12.7)
***A**nnoy* — Yes	71	38.0	(26.7 49.3)	23	39.1	(19.2 59.1)
***G**uilty* — No	872	20.0	(17.3 22.6)	947	9.5	(7.6 11.4)
***G**uilty* — Yes	86	34.9	(24.8 45.0)	49	46.9	(33.0 60.9)
***E**ye-opener* — No	948	21.7	(19.1 24.4)	992	11.3	(9.3 13.3)
***E**ye-opener* — Yes	13	7.7	(0.2 36.0)†	6	33.3	(11.8 77.7)†

† Exact Binomial Confidence Intervals. Other CIs from Normal approximation.

3.5 Summary

Information on alcohol consumption in the NSHD was collected using two instruments: weekly recall and seven-day diary. The weekly recall total was derived from one simple question asking for the total drinks in the previous week in the three categories of spirits, wine and beer. The seven-day diary involved recording all the food and drink consumed during each of the seven days of the week. The simpler recall question had a much higher response rate (97%) than the diary (61%).

However, the measurement of alcohol consumption based on the collection of information on drinking in diet diaries has a greater validity than that based on the recalled total. It was argued in Section 1.2.3 that the diary instrument is likely to give a more valid estimate of consumption than the recall. This chapter shows that the estimates of excessive drinking based on the simple weekly recall are significantly lower than those derived from the seven-day diet diary. Alcohol

consumption reported in the diary has greater validity than the recall for two reasons. First, the higher estimates given by the diary have greater credibility since all estimates from general population surveys underestimate alcohol consumption relative to national estimates based on revenue from sales. Second, subjects are unlikely to have had the motivation to exaggerate their drinking in the diary as have alcoholics in clinical treatment (personal communication, Professor John B. Davies, Centre for Applied Social Psychology, University of Strathclyde). The proportion of the sample who were classified in a higher drink category in the diary than in the recall measure is used as a measure of the extent to which the recall measure underestimated alcohol consumption. 'Under-reporting' is more prevalent in the recall than in the diary.

Although the recall measure underestimates consumption this could be adjusted for if the extent of under-reporting were similar for all respondents. However, the extent of under-reporting was associated with respondents' attitudes to drinking. Those who reported having problems with drinking in the CAGE questions were more likely to under-report their consumption in the recall. In addition, the extent of under-reporting was greater for men than for women but the association of under-reporting with attitudes to drinking was stronger for women than for men.

Using the diary to estimate the prevalence of excessive alcohol consumption poses the problem of how to cope with the large amount of missing data in the diary. Standard methods of analysis estimate the rate of excessive consumption using only those respondents who completed all the seven days of their diaries, that is only 61% of those who were interviewed. If those who completed the diary are not a random sample or representative in terms of their alcohol consumption of all those who were interviewed, then using only these completers will result in biased estimates. It cannot generally be assumed that those who do not complete questions are a random sample of all those interviewed in a survey. This is especially important when information is missing for a large proportion of respondents, as for the diet diary.

Missing data in surveys may be avoided by using simple instruments which minimise item non-response by requiring less commitment from respondents. However, in the case of alcohol consumption, it has been demonstrated that such simple instruments reduce the validity compared with the more complex instruments such as a detailed diet diary. The greater detail contained in a diary not only makes it more credible but also enables us to look at different aspects, or patterns, of alcohol consumption. The problem is that diet diaries, requiring considerable commitment on the part of the respondent, are poorly completed. The missing data poses a problem for analysis. Standard statistical methods can only use the completed records. Ignoring missing data by using complete cases is not only inefficient (since it discards incomplete information) but may result in bias. It is not possible to assess the bias directly since the missing values are not known, and there is no external source of information in this case.

Chapter 4

Dangers of Ignoring Item Non-Response

4.1 Introduction

This chapter demonstrates the bias in estimates that may result from ignoring incomplete records.

The previous chapter concluded that valuable information about alcohol consumption could be derived from the diet diary data. But a substantial fraction of respondents, 38.6%, did not complete the diet diary. Standard procedures for statistical analysis make use only of records with complete data. Such analyses exclude the respondents who did not complete the diary, using only those who completed the diary (*completers*). This not only decreases the efficiency of the analysis, because fewer cases are available, but also may give biased estimates. Furthermore, discarding partially complete records is a waste of the information contained in them.

If the problem of dealing with the missing data is ignored, then any analysis using alcohol consumption as a variable may be biased. It is not possible to identify any bias that results from ignoring missing data on alcohol consumption directly since the missing values are unknown. So we cannot identify the bias in estimates of alcohol consumption itself, or of its association with other variables, resulting from the use of cases with complete data only. Analyses in which alcohol consumption is a covariate give us the opportunity to identify any bias arising from the selection of completers only (i.e. cases with complete data in alcohol consumption). This situation arises in practice when alcohol consumption is included as a potential confounder. The example used here examines the association between birthweight and blood pressure in men in mid-life.

The "Barker hypothesis". Professor David Barker and colleagues have hypothesised that "a baby's nourishment... influences the diseases it will experience in later life." (Barker, 1994). Barker and colleagues have specifically examined the evidence for the early origins of heart disease. They conjectured that a baby's nourishment before birth and during infancy, as manifest in patterns of fetal and infant growth, "programmes" the development of risk factors for coronary heart disease in later life (Barker, 1992). These risk factors include fibrinogen concentration, factor VIII concentration, glucose intolerance and raised blood pressure (Paneth and Susser, 1995). The hypothesis has been widely challenged (for example, Ben-Shlomo and Davey Smith, 1991; Strachan et al., 1995; Christensen et al., 1995; Susser and Levin, 1999; Huxley et al., 2002), but remains the subject of study. Birthweight has been the most widely studied measure in retrospective studies, chiefly because of its availability from existing records

or personal recall. The NSHD provides the opportunity to examine, in a prospective study, the relationship between birthweight and adverse health outcomes in adulthood such as obesity (Kuh et al., 2002) or high blood pressure in mid-life (Wadsworth et al., 1985; Hardy et al., 2003). The association of adverse outcomes with lower birthweight seems to be strongest for blood pressure (Leon, 1999; Robinson, 2001) and for systolic rather than diastolic blood pressure (Huxley et al., 2002; Hardy et al., 2003).

One criticism of the Barker hypothesis is that the association could be due to confounding by socio-economic factors — either directly through socio-economic status or indirectly through health behaviours (Huxley et al., 2002). Childhood social circumstances could mediate the relationship between birthweight and health outcomes in later life. Low birthweight may be associated with low current social class (through childhood social class) and high blood pressure may also be associated with low social class through its relationship with poorer health behaviours. Alcohol consumption may be a confounder in the relationship between birthweight and blood pressure because of its relationship with blood pressure and social class. The positive association between alcohol consumption and blood pressure is well known (e.g. Bamford et al., 1990); and a stronger association with systolic than diastolic blood pressure has been reported (van Leer et al., 1994). Heavier drinking is traditionally associated with lower social class in men, though this has not been found in women (see Section 1.2.1).

Other lifestyle factors related to both high blood pressure and low social class include exercise and smoking. Another potential confounding factor is current BMI since it is positively associated with blood pressure and birthweight is positively associated with later size (Leon, 1998). Adjustment for BMI has been criticised because of difficulties of interpretation (Lucas et al., 1999) and for upwardly biasing relations between birthweight and blood pressure (Huxley et al., 2002). Of the 55 studies reviewed by Huxley and colleagues (2002), few had adjusted for potential confounding factors, such as parental socio-economic status (7), current socio-economic status (2), or alcohol consumption (3).

As we add covariates to an analysis, only cases with complete data are included, so any change in the coefficient of interest may arise because the estimate is sensitive to which cases are included (selection effect) or because of the confounding effect of the covariate.

4.2 Methods

Variables

Systolic blood pressure (SBP). Blood pressure was measured twice at 43 years by the nurse using a Hawksley random zero sphygmomanometer in mm Hg. As in previous analyses (Hardy et al., 2003), the second measure was used, unless this was missing or erroneous, when the first was used instead. The second measure is preferred, since the subject would be more relaxed.

Birthweight. Birthweight of cohort members, to the nearest quarter of a pound, was extracted from medical records within a few weeks of delivery and converted into kilograms.

Childhood social class was assigned from father's occupation. It was the social class of the father's occupation when the cohort member was 11 years of age or, if this was not available, at 15 years of age or, if this was not available, at 4 years of age (Kuh et al., 2002).

Current social status. Men were categorised as unemployed, employed in manual work or employed in non-manual work, using current or most recent occupation, based on the Director General's classification of social class of occupations. Manual work included those with occupations in social classes III manual, IV, V,and VI; non-manual included occupations in social classes I, II and III non-manual (OPCS, 1980).

Body Mass Index (BMI). Standing height and body weight were measured by the nurses, using a standard protocol recommended by the Royal College of Physicians (Williams et al., 1983), and BMI was calculated as weight (kg) divided by height (m) squared.

Exercise. Subjects were asked 'Do you regularly take part in any sports or vigorous leisure activities or do any exercise? (things like badminton, swimming, yoga, press-ups, dancing, football, mountain climbing or jogging)'. Responses were 'yes' or 'no'.

Smoking status. Subjects were asked 'Do you smoke cigarettes?' Responses were 'yes' or 'no'.

Alcohol consumption. The total alcohol consumed in the 7 day diet diary (measured in Units) was classified at 4 levels according to health criteria (see section 3.3), as shown in Table 4.1.

Table 4.1: Classification of levels of alcohol consumption

	Level of total alcohol consumption in Units	
Description	**Men**	**Women**
None	0	0
Sensible	over 0, up to 21	over 0, up to 14
Immoderate	over 21, up to 50	over 14, up to 35
Heavy	over 50	over 35

Analysis

First, the relationship is examined between blood pressure and alcohol consumption at age 43 in men and women who completed their diet diary (Section 4.3.1). Then the dependence of blood pressure on birthweight is examined (Section 4.3.2). The subsequent analyses refer to systolic blood pressure (SBP) for men only, as explained in Section 4.3.2.

The relationship between birthweight and SBP is summarised by the regression coefficient of SBP on birthweight. Potential confounders were considered in the following order: childhood social class, current social status; then current variables affecting blood pressure: body mass index (BMI), exercise, smoking status and alcohol consumption. Each potential confounder was added in turn as a covariate in the regression model. Smoking status was not included because it

did not add significantly to the model after adjusting for the other variables ($F_{1,1490} = 1.034$, $P = 0.309$). So six analyses are considered in turn: first unadjusted, second adjusting for childhood social class, third adjusting for childhood social class and current social class, and so on, finally adjusting for all the potential confounders, including alcohol consumption.

As each potential confounder is added, the number of cases available for analysis, i.e. those with complete data on all the variables included in the analysis, decreases (Section 4.3.3). For each set of cases, three regression estimates were evaluated: **A**, **B1** and **B2**, as described below.

A The unadjusted regression coefficient of blood pressure on birthweight (i.e. not including any other variables in the model) was estimated, using the subset of cases available for each of the six analyses.

B For each of the six analyses, using the corresponding subset of available cases, two adjusted coefficients were evaluated:

B1 Adjusting for all covariates from the preceding analysis and the current covariate.

B2 Adjusting for all covariates from the preceding analysis, but excluding the current covariate.

The models were fitted using SPSS (GLM procedure). In all analyses, the stratification of the birth cohort sample was taken into account by using a weighted combination of the estimates for each stratum as given in Section 2.6.1.

Any differences between the unadjusted coefficients for the different subsets of cases (as observed in **A**) can be ascribed to a selection effect. Any difference between the coefficients based on the same subset of cases (**B**), but using the model with (**B1**) and without (**B2**) a particular covariate can be ascribed to the confounding effect of that covariate on the estimated coefficient.

4.3 Results

4.3.1 The dependence of blood pressure on alcohol consumption in completers

The relationship between alcohol consumption and blood pressure is shown in Table 4.2 for the 961 men and 998 women who competed their diet diary and had measures of blood pressure. Relationships with alcohol consumption are similar for systolic (SBP) and diastolic blood pressure (DBP).

Table 4.2 shows that mean SBP and mean DBP increased with level of alcohol consumption in men. Mean SBP increased from 121.3 mm Hg in men who reported no alcohol consumption to 130.4 mm Hg in those who drank heavily. Similarly mean DBP in men increased from 79.0 to 84.6 mm Hg. On average, the women had lower blood pressure than the men (SBP 121.2 mm Hg in women, compared with 124.4 mm Hg in men; DBP 76.8 mm Hg in women and 81.5 mm Hg in men). However, women who drank heavily had considerably higher average blood

Table 4.2: Relationship between level of alcohol consumption and blood pressure (mm Hg)

Level of alcohol consumption	Men					Women				
		SBP		DBP			SBP		DBP	
	n	mean	sd	mean	sd	*n*	mean	sd	mean	sd
None	142	121.3	16.5	79.0	13.1	283	121.9	16.0	77.2	12.0
Sensible	500	123.4	14.2	81.5	11.2	564	120.1	15.6	76.0	11.4
Immoderate	238	126.5	15.7	82.0	12.0	142	123.2	16.7	78.3	11.1
Heavy	81	130.4	16.7	84.6	10.7	9	136.9	16.4	88.4	14.5
All	961	124.4	15.3	81.5	11.7	998	121.2	16.0	76.8	11.6

pressure (SBP 136.9 mm Hg, DBP 88.4 mm Hg) than those drinking less, even higher than men who drank heavily; but there were only 9 such women. Women who drank at sensible levels had the lowest mean blood pressures (SBP 120.1 mm Hg, DBP 76.0 mm Hg), those reporting no drinking having higher blood pressure on average (SBP 121.9, DBP 77.2) than those who drank, but within sensible limits. The relationship between alcohol consumption and both systolic and diastolic blood pressure is illustrated by the box and whisker plots for each level of drinking, shown in Figure 4.1. There is a monotone trend of increasing blood pressure with increasing alcohol consumption for men, while for women the relationship is 'U' shaped with those not drinking having higher blood pressure than those who drink sensibly.

4.3.2 The dependence of blood pressure on birthweight

Since the relationship between alcohol consumption and blood pressure is montone for men, but not for women, separate models would be needed for men and women. Results for women are not presented here. For the 1583 men for whom both blood pressure and birthweight was observed, SBP has a significant negative linear regression on birthweight (coefficient for linear term −2.35, se 0.760, $P = 0.002$), but a non-significant quadratic effect ($F_{1,1580} = 0.9078$, $P = 0.304$). However, DBP had no significant relationship with birthweight (coefficient for linear term −0.527, se 0.590, $P = 0.372$). Linear models were fitted for men only, with SBP as the continuous dependent variable.

4.3.3 The problem of missing data in the model for SBP in terms of birthweight

In general, the extent of missing data increases as more variables are included in the analysis. Table 4.3 gives the extent of missing data on the variables to be included in the model based on the 1635 men interviewed in 1989. Most of the variables are missing for a very small proportion of cases. For example, birthweight is missing for only 4 (0.2%) of the respondents. However, childhood social class is missing for almost 5% ($n = 78$) and alcohol consumption for 40.2% ($n = 657$) of the respondents. As we build the model for blood pressure in terms of birthweight, controlling for more potential confounders, the number of cases excluded because of incomplete records is cumulative. Of the 1635 respondents, SBP was measured for 1587, and 1583 also had their birthweight recorded. With the inclusion of childhood social class as a

Figure 4.1: Relationship between blood pressure and alcohol consumption

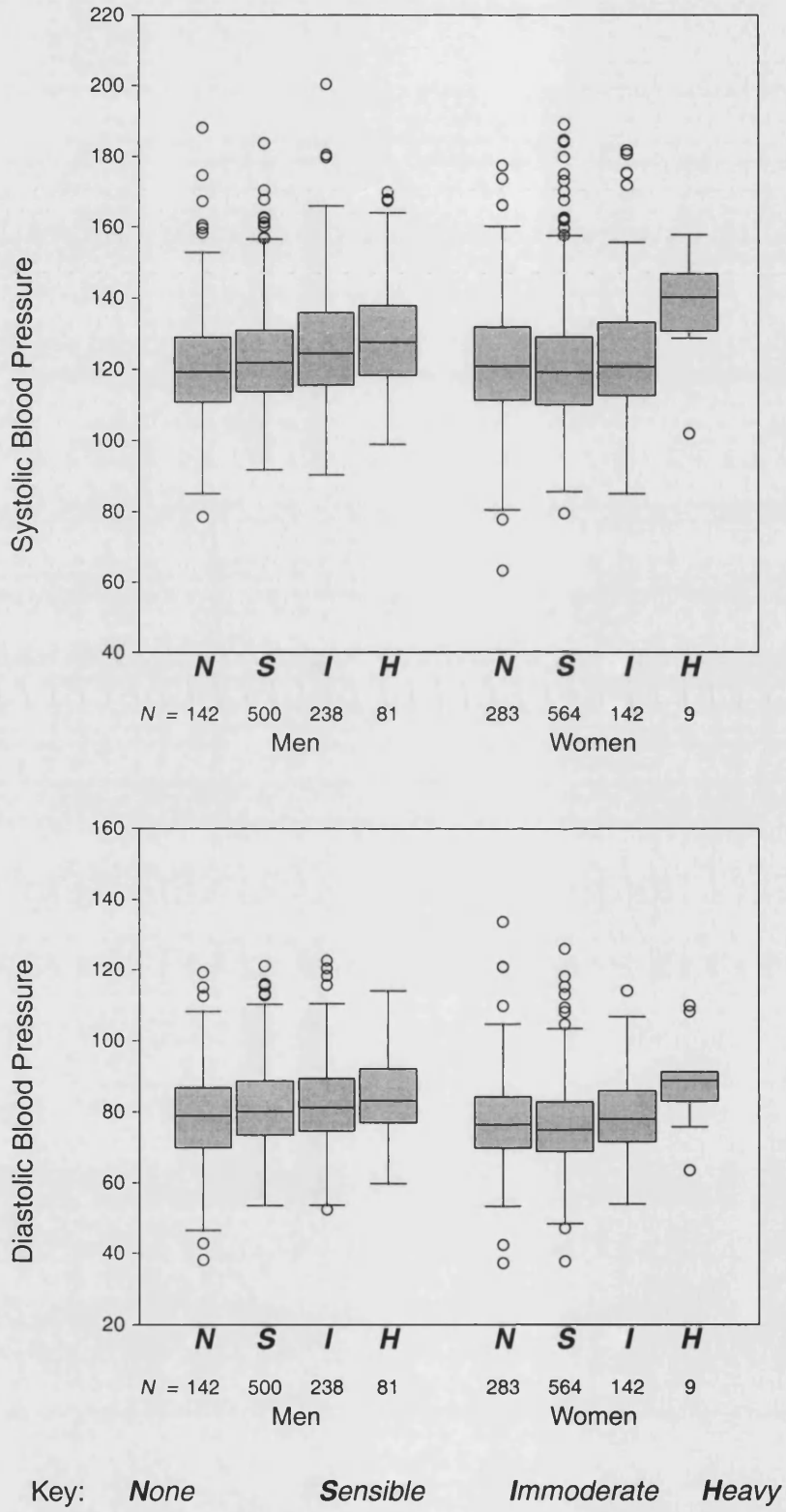


Table 4.3: Numbers of men with missing variables, and of cases available for analysis

				Cases available for the analysis		
Variable	*Number of cases missing*	*Percent of cases missing*	*Number of cases available*	*Cumulative number missing*	*Cumulative percent missing*	*Number available for analysis*
Men	0	0.0	1635	0	0.0	1635
SBP	48	2.9	1587	48	2.9	1587
Birthweight	4	0.2	1631	52	3.2	1583
Childhood social class	78	4.8	1557	128	7.8	1507
Current social status	3	0.2	1632	130	8.0	1505
BMI	19	1.2	1616	132	8.1	1503
Exercise	9	0.6	1626	136	8.3	1499
Alcohol consumption	657	40.2	978	723	44.2	912

covariate only 1507 cases are available for the analysis, 128 (7.8%) having missing data on at least one of the variables SBP, birthweight or childhood social class. Finally, when alcohol consumption is added in the model, only 912 cases are available for the analysis, 723 (44.2%) having missing data on at least one of the variables involved in the final model.

4.3.4 The coefficient in the regression of systolic blood pressure on birthweight

As we add the covariates to the regression model, only cases with complete data are included, so any change in the coefficient of birthweight may arise either because the estimate is sensitive to which cases are included or because of the confounding effect of the covariate. In order to detect a selection effect as each variable is successively added to the model, the regression coefficient of birthweight is estimated using the cases available for each analysis, but without adjusting for the effect of these variables (i.e. they are not included in the regression model) (**A** in Section 4.2). These results are shown in Table 4.4.

Table 4.4: Unadjusted coefficient in regression of systolic blood pressure on birthweight (mm Hg/kg) for men, using the subset of cases available for each analysis (A)

Covariates	***n***	**coeff**	**se**	**95% CI**	***P***
None	1583	−1.65	0.89	(−3.39, 0.08)	0.062
Childhood social class	1507	−1.57	0.91	(−3.36, 0.22)	0.086
Current social status	1505	−1.51	0.92	(−3.31, 0.28)	0.098
BMI	1503	−1.50	0.92	(−3.30, 0.29)	0.102
Exercise	1499	−1.53	0.92	(−3.34, 0.27)	0.096
Alcohol consumption	912	−3.50	1.22	(−5.89,−1.12)	0.004

A change in the coefficient of birthweight indicates its sensitivity to the cases included in the analysis. The coefficient based on the 1583 respondents with SBP and birthweight is −1.65, indicating that a 1 kg increase in birthweight was associated with a 1.65 mm Hg decrease in blood pressure. The coefficient changes little as successive variables are added to the list,

except for alcohol, which is available only for diary completers. The coefficient estimated for the 912 cases with complete data on all variables including alcohol consumption (–3.50 mm Hg/kg), however, is quite different from the others. The increase in the standard errors for the estimates reflects the lower efficiency due to the reduction in the number of cases available for each analysis.

As covariates are successively added to the model, and thus controlled for, the estimated coefficient of birthweight may change. In order to see whether the changes in the estimated birthweight coefficient are due to a confounding effect of the covariate being added, two analyses are compared (**B1** and **B2** in Section 4.2). The two analyses use the same set of cases, but in the second analysis (**B2**) the current variable is not controlled for. Hence the two analyses differ only in whether or not the current covariate is controlled for. Any difference in the coefficient between the two analyses can be ascribed to the confounding effect of the current variable.

Table 4.5: Adjusted coefficient in the regression of systolic blood pressure on birthweight (mm Hg/kg) for men (B — analyses B1 and B2)

		B1				B2			
Covariates	***n***	**coeff**	**se**	**95% CI**	***P***	**coeff**	**se**	**95% CI**	***P***
None	1583	–1.65	0.89	(–3.39, 0.08)	0.062				
Childhood social class	1507	–1.55	0.91	(–3.33, 0.24)	0.090	–1.57	0.91	(–3.36, 0.22)	0.086
Current social status	1505	–1.48	0.91	(–3.27, 0.31)	0.105	–1.49	0.91	(–3.28, 0.30)	0.103
BMI	1503	–1.66	0.91	(–3.45, 0.13)	0.069	–1.47	0.91	(–3.26, 0.32)	0.108
Exercise	1499	–1.72	0.91	(–3.50, 0.07)	0.060	–1.70	0.92	(–3.50, 0.09)	0.064
Alcohol consumption	912	–3.79	1.19	(–6.13,–1.45)	0.002	–3.76	1.19	(–6.10,–1.43)	0.002
		Adjusting for all covariates from the preceding analysis and also the current covariate				*Adjusting for all covariates from the preceding analysis, excluding the current covariate*			

Each row of Table 4.5 gives the estimated coefficient from two analyses on the same set of cases. The analysis on the left controls for all the covariates down to the one in the current row, the one on the right controls for all covariates in previous rows but not the covariate in the current row. The corresponding coefficients for the two analyses (**B1** and **B2**) differ very little with the exception of the row where BMI is introduced. Using the 1503 cases with complete data on childhood social class, current social status and BMI, controlling for childhood social class and current social status gives an estimate of –1.47 mm Hg/kg. When BMI is controlled for in addition, the estimate of the birthweight coefficient is –1.66 mm Hg/kg. The second estimate is more negative as a result of controlling for BMI, indicating a degree of confounding by BMI of the relationship between birthweight and SBP. Analysis of the 912 cases with data on all the variables, controlling for all the covariates, including alcohol consumption, gives an estimate of –3.79 mm Hg/kg. Controlling for all covariates except alcohol consumption gives a

similar estimate of −3.76 mm Hg/kg, indicating that alcohol consumption is not a confounder of the relationship between birthweight and SBP.

4.3.5 Discussion

The unadjusted estimate of the regression coefficient of systolic blood pressure on birthweight, −1.65 mm Hg/kg, is similar to that found by Huxley and colleagues (2002) based on similar sized studies. The finding that adjustment for BMI increases the negativity of the coefficient (by 0.2 mm Hg) was also found in Huxley's review. It is believed to indicate that it is change in relative size between birth and later life (low birthweight babies who later had a high BMI) that is implicated in later health outcomes (Lucas et al., 1999).

The similarity of the coefficients in the analyses adjusted for alcohol consumption (−3.79 mm Hg/kg) compared to that using the same cases but without adjusting for alcohol consumption (−3.76 mm Hg/kg) indicates that this covariate acts independently on the relationship between birthweight and systolic blood pressure. The change in the regression coefficient of birthweight on systolic blood pressure when alcohol consumption is controlled for (from −1.72 to −3.79 mm Hg/kg) is not due to confounding by alcohol consumption but to the selection effect of including only cases with complete diet diaries. The relationship between birthweight and systolic blood pressure is not the same for those who do not complete their diet diary as for those who do. Low birthweight may be a risk factor for high blood pressure in men in mid-life, but the risk is apparently much greater in those respondents who completed their diet diaries. Ignoring missing data for alcohol consumption results in lower efficiency because of the reduction in the number of cases available for the analysis (the standard error increases from 0.91 to 1.19), and a substantially overestimated association between birthweight and systolic blood pressure in men.

4.3.6 Summary

In epidemiological analyses of longitudinal data, the problem of missing data is cumulative as the number of variables included in the analysis increases. When there are missing values in a covariate to be included in an analysis, the missing data may be considered as just a nuisance. Since standard methods use only cases with complete data, as a covariate is added to a model, some cases are excluded because they have missing values for this covariate. If no consideration is given to the effect of missing data, the effect of omitting cases is confounded with the modifying effect of the covariate on the estimate of interest. We cannot tell whether a change in the estimate is due a selection effect or to the modifying effect of the covariate on the estimate.

The bias that may result from ignoring missing data cannot be demonstrated directly since the missing values are by definition unknown. However, the danger of bias has been demonstrated when the missing variable is used as a covariate in an analysis. Even if those who fail to complete their diet diaries are not different in their drinking habits from those who complete their diaries, they may be different in other ways. Failure to deal with missing data can bias the results of any analysis using a variable with missing values, besides reducing the efficiency of the analysis by reducing the number of cases available. The results of any such analysis can be

seriously misleading. Hence it is important to find a method for dealing with missing data that allows the analyst to make use of cases with incomplete records, whatever the role of the variable with missing data in the analysis. A method for dealing with missing data should make use of all the available information. The next Chapter investigates the available information about non-response and about alcohol consumption in the diet diary.

Chapter 5

Factors Associated with Non-Response and with Alcohol Consumption in the Diet Diary

5.1 Introduction

Chapter 3 demonstrated that valuable information about alcohol consumption could be derived from the diet diary data, but a large number of NSHD respondents did not complete the diet diary. Using conventional methods of epidemiological analysis, the cases with incomplete records are excluded. Estimates based on such analyses are inefficient, and biased unless the data is missing completely at random. Chapter 4 demonstrated that bias results from ignoring missing data on alcohol consumption. It found that the fitted association between birthweight and blood pressure amongst male completers of the diary substantially overestimates the association in the target population. .

Estimates of alcohol consumption itself are biased if those who did not complete their diaries drank more (or less) than those who do. We cannot tell whether those who did not complete their diary actually consumed more alcohol than those who did, unless it is possible to collect the missing information in a follow up. As this is usually not possible, a method used in epidemiology is to ascertain indirectly whether there are differences between respondents with complete and with incomplete diary data, and whether these differences are associated with alcohol consumption. The results in Chapter 4 demonstrate that there are differences between these two groups of respondents, so the diary data are not missing completely at random (MCAR). In general, non-response is unlikely to be a purely random process, and we have only to find an observed characteristic that is related to non-response to confirm this. However, this need not pose a problem for methods of dealing with the missing data that take into account what we know about alcohol consumption, provided the data is missing at random (or MAR, see Section 2.5.1.) conditional on the variables associated with alcohol consumption. Imputation provides such a method for dealing with missing data (Section 2.7.2). An imputation model should include the variables that relate most directly to alcohol consumption and also those that are associated with non-response.

This chapter examines the information available in the NSHD about those who did not complete their diet dairy and about alcohol consumption reported in the diary, in order to develop the model for imputation. We start by reviewing the evidence in the literature about the factors associated with non-response and with alcohol consumption.

5.1.1 Factors associated with non-response

It is generally reported that non-response in sample surveys is greater for men than women, for those in low social classes compared to higher social classes, and for those with lower educational qualifications (Goyder, 1987). This is found in different countries and for different types of survey. For example, the association of non-response with male gender has been reported in a self-completion lifestyle questionnaire conducted in the UK (Dengler et al., 1997), in a health survey in Finland (Korkeila et al., 2001) and in a questionnaire survey of a twincohort in Australia (Heath et al., 2001). The association of non-response with low social status has been reported in the UK (Dengler et al., 1997) and in a mail survey of disabled people in the US (Sheikh & Mattingly, 1981), or with low income in a lifestyle assessment in an elderly cohort in the US (Slymen et al., 1994). Of these studies, all those that included information about educational qualifications found non-response to be associated with lower educational qualifications (Heath et al., 2001; Dengler et al., 1997; Slymen et al., 1994). Non-response is also reported as being associated with unemployment (Sheikh, 1986) and being divorced or widowed (Korkeila et al., 2001) or being unmarried (Slymen et al., 1994).

5.1.2 Factors associated with alcohol consumption

There is an extensive literature describing the associations of various factors with alcohol consumption. This research generally focuses on identifying risk factors for very high levels of alcohol consumption or alcoholism. There is no generally agreed way to measure alcohol consumption (see Section 1.2.3); the literature differs not only in the methods used to collect the data (such as interview vs. self-completion, quantity-frequency measures vs. daily diary) but also in the definition of high alcohol consumption. Much lower levels of drinking are consistently reported in women compared with men (Plant, 1997). Another factor is the decline in consumption with age in adulthood (Johnson et al., 1998). Comparisons between studies can therefore be made only after adjusting for age, but age adjusted results are often not reported. When lower levels of drinking are included, the relationship between the factor and alcohol consumption is often not monotone. For example, both abstaining and excessive drinking are associated with lower education in men (Greenfield et al, 2000; van Oers, 1999), and also with unemployment (Lee et al., 1990). The relationship with education is not so clear: some report little relationship (Cactano & Clark, 2000) others report association of heavier drinking with more educational qualifications (Knibbe et al, 1985). Getting married or becoming a parent is associated with drinking less (Hajema & Knibbe, 1998); but being divorced or separated is associated with drinking more (Power et al, 1999). Smoking is positively associated with drinking (Istvan & Matarazzo, 1984; Launer et al., 1996). The positive association between alcohol consumption and blood pressure is discussed in Section 4.1.

Patterns of drinking over the days of the week

The day of the week is a predictor of alcohol consumption on any particular day of the week. In the past, the pattern of drinking over the days of the week has been observed to vary with occupational social class. This pattern is typified by Londoners in the 1960s (Edwards et al.,

1972). Edwards' work supports the stereotype of the working class man bingeing in the pub at the weekend, while the white-collar worker takes his drink more regularly and moderately over the week, perhaps as wine with meals. In Edward's study the women in higher social classes drank more than those in lower social classes. This is also a characteristic of the OPCS drinking surveys in the 1980s, and in other studies both in Britain and other European countries (Makela, 1999).

5.2 Socio-economic factors associated with non-response to the diet diary.

We now examine the association of factors with having a complete diary, i.e. when all seven days of the diary are completed, compared with an incomplete diary, i.e. fewer than seven days completed. Statistical tests, using χ^2 for categorical variables and t-tests for continuous variables, are applied to those with non-missing values on these variables. The proportion with incomplete diaries is reported for those with missing data on the variables. Adult social class is based on current occupation coded using the Registrar General's Classification of Occupations (OPCS, 1980). If the subject was not employed at 43, the most recently available occupation was used based on information recorded at the age of 36 or 26. Sixty cases (13 men and 47 women) did not have a social class defined in this way: 42 women have their spouse's social class, of the remaining 18, 16 have a social class based on their parental occupation during childhood, and two cases for whom none of this information is available have their parental social class at birth.

The association of socio-economic factors with missingness of the diet diary is summarised in Table 5.1, since these factors are known to be associated with non-response (Section 5.1.1). Those from a manual social class were more likely to have an incomplete diary. This applied to both adult social class (manual 42.1% incomplete vs. non-manual 36.9% incomplete) and to social class of origin as measured by the father's social class at birth, although manual workers here excludes agricultural workers (manual (not agricultural) 41.9% vs. non-manual and agricultural 36.1%). Men were somewhat more likely to have incomplete diaries than women (40.2% vs. 36.9%). Employment status was not indicative of completion of the diary; although differences were somewhat greater amongst men than women they were not statistically significant at the 5% level (results by gender are not presented). In terms of marital status, those who had never married were most diligent (only 33.5% failed to complete), whilst those who had been married but were currently widowed, divorced or separated were least likely to complete their diaries. Similar relationships held amongst men and women in respect of marital status. The expected negative relationship between non-response and educational qualifications attained by the age of 26 years showed non-response increasing with lower educational qualifications. Poor educational attainment may make the completion of a complex instrument like the diet diary more demanding. Self-completion of a survey questionnaire requires the ability to read, and, when the answers are not simply tick boxes, the ability to write. Some

Table 5.1: Socio-economic factors associated with missing data in diet diaries (all respondents, $N = 3262$)

Factor	N	% with incomplete diaries	χ^2	df	*P*
Father's social class at birth					
Non-manual and agricultural workers	1859	36.1			
Manual (but not agricultural) workers	1403	41.9	11.20	1	0.001
Adult social class					
Non-manual	2171	36.9			
Manual	1091	42.1	8.21	1	0.004
Gender					
Female	1627	37.1			
Male	1635	40.2	3.35	1	0.07
Employment status					
Employed	2846	38.4			
Non-employed	318	41.2			
Unemployed	97	36.1	1.201	2	0.546
Missing	1				
Marital status					
Never married	218	33.5			
Married	2606	38.0			
Widowed	37	45.9			
Separated	86	45.3			
Divorced	315	44.8	10.35	4	0.035
Educational qualifications at 26 years					
Degree	300	26.3			
Vocational to A level	1634	37.4			
None	1141	42.0			
Handicapped	67	49.3	28.70	3	<0.001
Unknown	120	48.3			
Difficulty with reading					
No	3166	38.3			
Yes	88	47.7	2.84	1	0.092
Missing	8	75.0			
Difficulty with writing or spelling					
No	2986	37.6			
Yes	267	49.4	14.56	1	<0.001
Missing	9	66.7			
Difficulty with sums and calculations					
No	3109	38.1			
Yes	143	46.2	3.37	1	0.066
Missing	10	80.0			

people may simply not have the skills required and those with a poor level of skills may be too embarrassed or lack confidence to attempt the task. This is particularly true of the diet diary which demands a great deal of writing, as well as some degree of numeracy to measure and record quantities. In the survey instrument at the age of 43, respondents were asked whether they had difficulty with basic reading, writing, or arithmetic. The relationship with completion of the diary at the time can be expected to be more direct than that of educational qualifications.

This is substantiated by the proportion of people with difficulties in these areas who do not complete their diaries. In particular of the 267 who reported difficulties with writing or spelling, almost half of them (49.4%) failed to complete their diaries.

A number of socio-economic factors are strongly related to completing the diary, in particular, the current social class, but also the father's social class at birth, on which the birth cohort was stratified, and education and basic skills. Clearly, the diet diaries, and hence the diary data on alcohol consumption, are not missing completely at random. The next section examines the information that we have about alcohol consumption in the diet diary for those who complete their diaries.

5.3 Factors associated with alcohol consumption in the diet diary

We now examine the association of factors with alcohol consumption reported in the diet diary by the 2002 NSHD survey members at the age of 43 who completed their diary. Alcohol consumption is measured as the total alcohol consumed in the seven-day diet diary classified in four levels (measured in Units), as given in Table 4.1.

There is a large difference between alcohol consumption levels of men and women (Plant, 1997). Biological evidence suggests that women are more sensitive to the physiological effects of alcohol than men and therefore studies of alcohol problems should allow for a gender-specific level of alcohol consumption (Ely et al., 1999; Plant, 1997). This biological difference is reflected in the generally accepted recommended safe drinking levels (Royal College of Psychiatrists, 1986). Table 5.2 shows that even with gender-specific levels of consumption, smaller proportions of women than men drink either immoderately (14.6% vs. 24.6%) or heavily (1.1% vs. 8.6%). In particular, women are more likely to report not drinking at all (28.1% women vs. 14.8% men).

Table 5.2: Alcohol consumption of men and women in the diet diary
(Completers only: $N = 2002$)

	Range of total alcohol consumption, Units							
	MEN (N= 978)				WOMEN (N=1024)			
	0	**0–21**	**21–50**	**>50**	**0**	**0–14**	**14–35**	**>35**
	none	**sensible**	**immoderate**	**heavy**	**none**	**sensible**	**immoderate**	**heavy**
n	145	508	241	84	288	576	149	11
%	14.8	51.9	24.6	8.6	28.1	56.3	14.6	1.1

The gender-specific levels of assessing drinking and the large gender differences in distribution of alcohol consumption imply that any associations should be considered separately for men and women.

Chapter 3 discussed two other measures recorded at the 43 year interview, besides the seven-day diary, which directly relate to the alcohol consumption of the respondents: the amount of alcohol consumed in the previous week (weekly recall), and the CAGE score. The total alcohol consumed in the seven-day dairy is strongly associated with the weekly recall (Section 3.4.1), and weekly recall to the CAGE score (Ely et al., 1999). The CAGE score influences the relationship between the two measures of alcohol consumption (Section 3.4.2). Table 5.3 describes the level of alcohol consumption in relation to these and other factors that are correlates of alcohol consumption in the diet diary.

Although the weekly recall underestimates consumption relative to the diary, as discussed in Section 3.4.1, the strong correlation between the two measures can be observed in the high proportion of people on the main diagonals of the tabulation of weekly recall against seven-day diary levels (printed in boldface in Table 5.3). The CAGE score is considered at two levels, with scores of 2, 3 and 4 considered to be indicative of problem drinking, and scores of 0 and 1 otherwise. Both men and women who report problem drinking (CAGE score 2-4) drink considerably more on average than those who do not (CAGE score 0 or 1). For example, men who report problem drinking are more than four times as likely to drink heavily (26.7%) as those who do not (6.5%). For women, the differences are even greater (20.8% vs. 0.1%). Those with drink problems are also much less likely to report no drinking in the diary week than others (3.8% vs. 16.1% for men; 4.2% vs. 28.9% for women). Similar, though less extreme, differences exist amongst men who smoke compared with non-smokers. Differences for women are in the same direction: smokers drink more than non-smokers. This is mainly because women smokers are more likely to drink excessively (above 14 Units) than non smokers, and not that smoking women are any more likely to drink at all (28.5% of women smokers do not drink in the diary week compared with 28.0% of non-smokers). Both men and women in manual occupations are more likely not to drink than those in non-manual occupations (17.4% vs. 13.3% of men and 39.5% vs. 23.9% of women), but for men being in a manual occupation is also associated with heavy drinking (11.8% vs. 6.8%), whereas women in non-manual occupation are more likely to drink excessively (above 14 Units) than those in manual occupations (a total of 17.7% (16.2 + 1.5) compared with 10.1% (10.0 + 0.1)). The relationship between marital status and alcohol consumption is not straightforward. Married subjects tend to drink more moderately than those who are not currently married, but are less likely to report no drinking. Those who are separated or divorced tend to drink less moderately but men with this status are also more likely to report no drinking than those who were married.

The dependence of blood pressure on alcohol consumption in completers is discussed in Section 4.3.1.

Table 5.3: Factors associated with alcohol consumption in the diet diary: Percentages of men and women drinking nothing, sensibly, immoderately and heavily. (Completers only: *N* = 2002)

		MEN (% of *N*) Range of total alcohol consumption, Units					WOMEN (% of *N*) Range of total alcohol consumption, Units			
Factor		**0**	**0–21**	**21–50**	**>50**		**0**	**0–14**	**14–35**	**>35**
Weekly recall (M/W)	*N*					*N*				
0	148	**63.5**	36.5	0	0	300	**71.3**	28.7	0	0
0–21/0–14	636	7.7	**66.8**	23.6	1.9	636	9.0	**74.2**	16.5	0.3
21–50/14–35	151	0	13.2	**57.0**	29.8	58	3.4	15.5	**69.0**	12.1
50+/35+	29	0	0	6.9	**93.1**	4	0	0	50.0	**50.0**
Missing	14	14.3	64.3	21.4	0	26	51.7	34.6	0	7.7
CAGE score	*N*					*N*				
2–4 (drink problem)	105	3.8	24.8	44.8	26.7	48	4.2	29.2	45.8	20.8
0–1 (no drink problem)	862	16.1	55.2	22.2	6.5	961	28.9	58.0	13.0	0.1
Missing	11	18.2	54.5	27.3	0	15	53.3	33.3	13.3	0
Smoker	*N*					*N*				
Yes	267	11.6	39.7	32.6	16.1	263	28.5	49.8	19.4	2.3
No	710	16.1	56.6	21.7	5.6	761	28.0	58.5	12.9	0.7
Missing	1				100.0	0				
Social class	*N*					*N*				
Manual	356	17.4	46.1	24.7	11.8	276	39.5	50.4	10.1	0
Non-manual	622	13.3	55.3	24.6	6.8	748	23.9	58.4	16.2	1.5
Marital status	*N*					*N*				
Never married	88	28.4	43.2	19.3	9.1	57	36.8	47.4	12.3	3.5
Married	792	12.9	54.4	25.3	7.4	824	27.5	58.1	13.5	0.8
Widowed	4	25.0	25.0	50.0	0	16	37.5	50.0	12.5	0
Separated	20	20.0	50.0	15.0	15.0	27	22.2	59.3	11.1	7.4
Divorced	74	17.6	37.8	25.7	18.9	100	28.0	46.0	26.0	0
Employment status	*N*					*N*				
Employed	925	13.6	52.5	25.4	8.4	828	25.8	57.9	15.6	0.7
Non-employed	31	41.9	41.9	3.2	12.9	156	39.7	49.4	9.0	1.9
Unemployed	22	27.3	40.9	22.7	9.1	40	30.0	50.0	15.0	5.0
Education at 26 years	*N*					*N*				
Degree	164	11.0	62.2	23.2	3.7	57	22.8	57.9	14.0	5.3
Vocational to AL	447	13.4	57.9	25.1	9.6	576	23.6	59.0	16.1	1.2
None	322	18.6	48.4	23.3	9.6	340	36.5	51.2	12.1	0.3
Handicapped	17	23.5	35.3	29.4	11.8	17	35.3	41.2	23.5	0
Unknown	28	10.7	42.9	39.3	7.1	34	26.5	64.7	8.0	0

5.4 Association with non-response of the factors related to alcohol consumption

This section investigates whether the factors associated with alcohol consumption are also associated with non-response in the diary. The most important factors to include in a model for missing alcohol consumption data are those that are associated with alcohol consumption and with non-response. Table 5.4 shows how the factors that are associated with alcohol consumption in the diet diary (and which are not included in Table 5.1) are associated with the non-response to the diet diary. Similar relationships were found for men and for women, so the results are presented only for all subjects together.

Table 5.4: Association of drink related variables with missing data in diet diaries (all respondents, $N = 3262$)

Factor	N	% with incomplete diaries	χ^2	d.f.	P
Weekly recall					
0	692	35.3			
0–14/0–21	2061	38.3			
14–35/21–50	351	40.5			
35+/50+	71	53.5			
Missing	87	54.0	19.25	4	0.001
CAGE score					
2–4 (drink problem)	265	42.3			
0–1 (no drink problem)	2929	37.8			
Missing	68	61.8	17.76	2	<0.001
Smoker					
Yes	2282	45.5			
No	973	35.5			
Missing	7	87.5	35.28	2	<0.001

	N	Mean difference between those with incomplete and complete diaries (mm Hg)†	t	df	P
Diastolic blood pressure	3157	1.39	3.008	3155	0.003
Systolic blood pressure	3157	1.17	2.000	3155	0.046

†(*Incomplete – Complete*)

Table 5.4 shows that the higher the level of drinking reported in the weekly recall, the more likely are subjects not to complete their diet diary. Of those reporting no drinking in the weekly

recall, 35.3% did not complete their diaries, a proportion that rose to 53.5% of those drinking heavily. Those reporting problem drinking (CAGE scores 2, 3 or 4) are more likely not to complete their diaries: 42.3% had incomplete diaries compared to 37.8% of those who did not report problems with drinking. Smokers were also more likely to fail to complete their diary (45.5% vs. 35.5% for non-smokers). Those with incomplete diaries had higher blood pressure than those who completed their diaries. The difference was greater for diastolic blood pressure (by 1.39 mmHg) than for systolic blood pressure (by 1.17 mmHg).

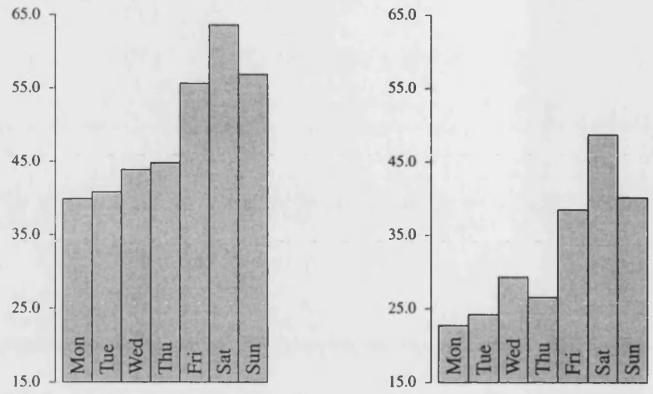
Each of these factors is positively associated with both higher alcohol consumption and with having an incomplete diary. It seems that those who were likely to drink more were also more likely to have incomplete data. However, of the 1260 subjects with incomplete diaries only 97 have no diary days at all (Table 3.1) and for 92.3% (1163/1260) at least two days of their diaries are available. Knowing what people drank on these two days gives us valuable information about their drinking on subsequent days. However, in general, people do not drink a similar amount on each day — the pattern of drinking varies with the day of the week.

5.5 Patterns of drinking over the days of the week

Alcohol consumption has a semicontinuous distribution (Section 2.4) that consists of two parts: a set of zeros (people who do not drink) and a separate continuous log Normal distribution (the log of the positive amount for those who did drink). These two parts of the distribution correspond to two aspects of drinking: whether someone drinks at all (the sign of drinking) and, if they do drink, how much they drink (positive amount of alcohol consumed). The pattern of drinking over the week is therefore considered from the two aspects: the proportion of people drinking at all (with positive sign of drinking) and the mean of the logged positive amount of alcohol consumed (amount).

Figure 5.1 illustrates these patterns for men and women. On the graph of the amounts each increment of 0.2 on the vertical axis represents a multiplicative increase of 22%. For example, the geometric mean of alcohol consumed for men on Saturdays is 5.0 Units (39.4 gm = exponential of 3.67), and they drank on average 42% more on Saturdays than Mondays (geometric mean for Monday is 3.5 Units, 27.8 gm = exponential of 3.33). Women drank on average 48% more on Saturdays (geometric mean 2.5 Units, 20.1 gm = exponential of 3.00) than on Mondays (geometric mean is 1.7 Units, 13.5 gm = exponential of 2.61). Figure 5.1 shows that fewer women than men drank, and they tended to drink lower amounts than men when they did drink. However the patterns over the days of the week are striking and similar for men and women. The proportion of people drinking (signs) depends on the day of the week. Saturday is the most popular drinking day, followed by Sunday and Friday, and these three days are referred to as the 'weekend'. Lower proportions of people drank on the weekdays (Monday to Thursday) than at the weekend, and these proportions varied less than they did at the weekend. As well as the popularity of drinking (how many people drink), the amount people drank also varied over the days of the week, and did so a similar pattern. The more popular the

Signs: proportion (%) consuming any alcohol



Amounts: mean of logged positive quantity (in gm) of alcohol consumed

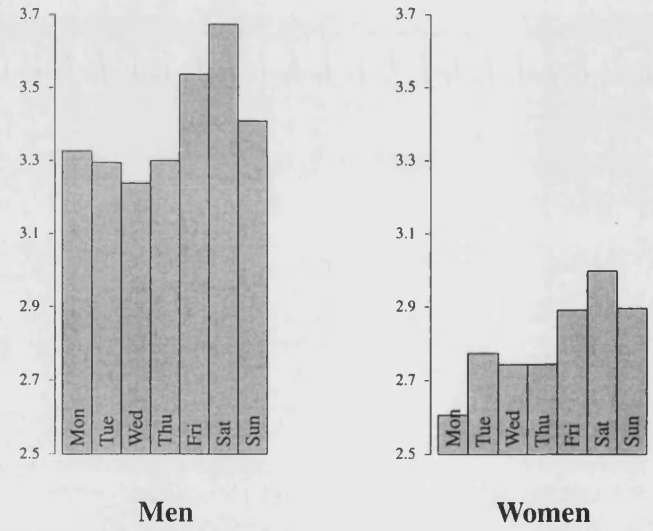


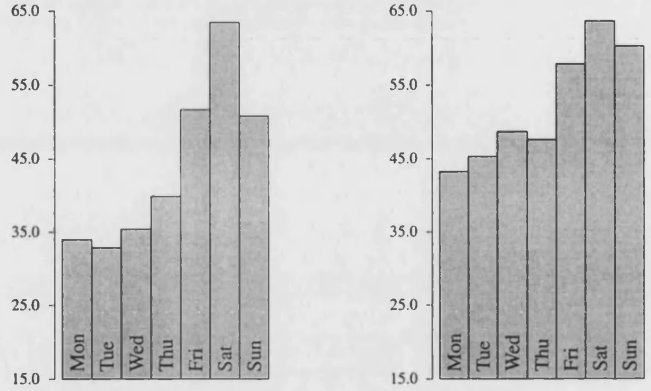
Figure 5.1: Patterns of drinking over the days of the week for men and women

drinking day, i.e. the greater the proportion of people who drank on a particular day of the week, the more alcohol was consumed by those who did drink. Drinking is a social activity.

Figures 5.2 and 5.3 give the drinking patterns for men and women (respectively), comparing those in manual and in non-manual occupations. Men in non-manual occupations were more likely to drink than those in manual occupations on any day of the week except Saturday, but when they drank they did so more moderately on average (Figure 5.2). Women in non-manual occupations were much more likely to drink than those in manual occupations on any day of the week, and when they drank they consumed just as much on average (Figure 5.3). So, men in higher social classes drank more often but more moderately than men in lower social classes;

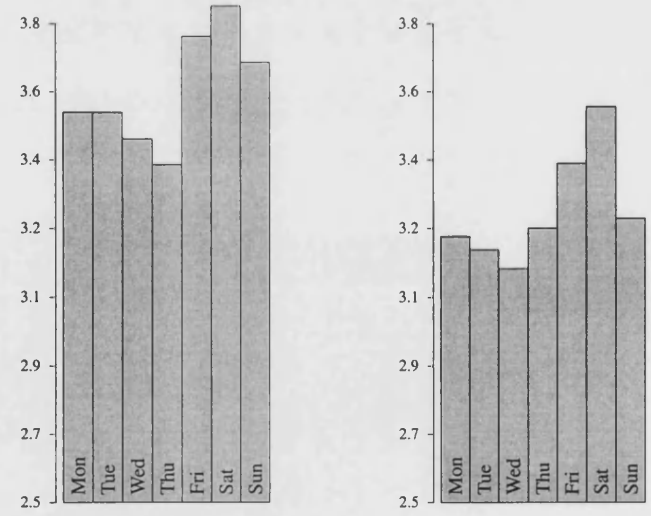
Figure 5.2: Patterns of drinking over the days of the week for men in manual and non-manual occupations

Signs: proportion (%) consuming any alcohol



Manual **Non-manual**

Amounts: mean of logged positive quantity of alcohol consumed

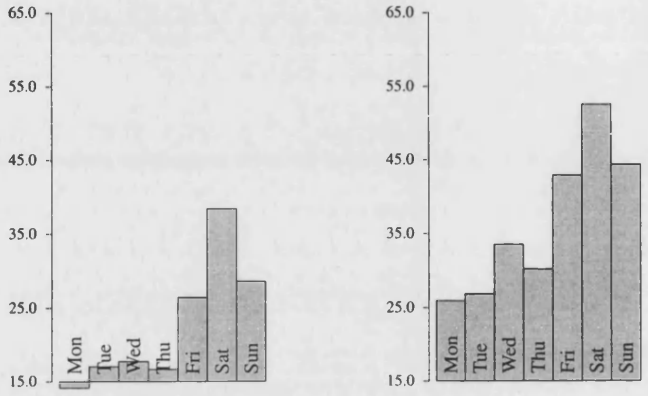


Manual **Non-manual**

MEN

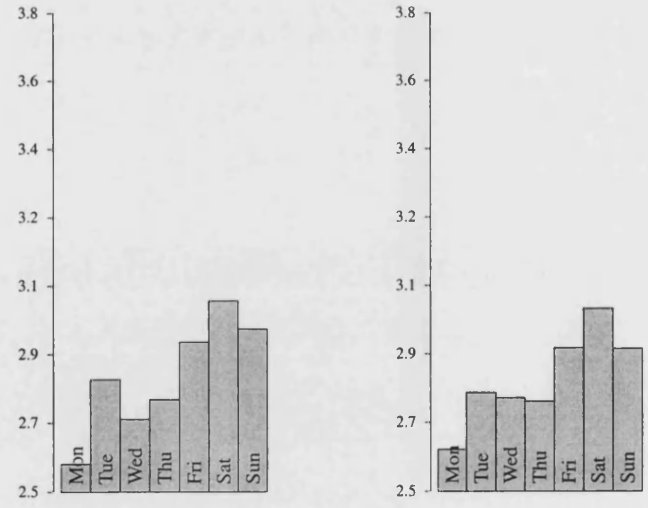
Figure 5.3: Patterns of drinking over the days of the week for women in manual and non-manual occupations

Signs: proportion (%) consuming any alcohol



Manual **Non-manual**

Amounts: mean of logged positive quantity of alcohol consumed



Manual **Non-manual**

WOMEN

but women in higher social classes drank more often than those in lower social classes, although they do not moderate their level of drinking.

The pattern of drinking over the week, being higher during weekends than during the week, is evident in all subgroups. However, the differences between the manual and non-manual social classes are not the same for men and women, implying an interaction between gender and social class in relation to pattern of drinking over the week.

Besides the effect of the days of the week on the average level of drinking, subjects differ in their patterns of drinking over the week. The more often an individual drank the more alcohol they consumed on average each day (Table 5.5).

Table 5.5: Mean alcohol consumption per day (mean of logged average positive alcohol consumption, in gm, per drinking day) by number of drinking days (Completers only: $N = 2002$)

		Number of drinking days							
		1	**2**	**3**	**4**	**5**	**6**	**7**	**1–7 days**
Men	**n**	106	112	141	128	103	103	140	833
	%	12.7	13.4	16.9	15.4	12.4	12.4	16.8	100.0
	mean	2.92	3.19	3.50	3.45	3.56	3.56	3.81	3.45
Women	**n**	188	149	107	89	78	71	54	736
	%	25.5	20.2	14.5	12.1	10.6	9.6	7.3	100.0
	mean	2.57	2.75	2.85	2.94	3.09	3.08	3.18	2.84
All	**mean**	2.70	2.94	3.22	3.24	3.36	3.36	3.64	3.16

The average alcohol consumption per day is calculated as the total alcohol consumption during the week divided by the number of drinking days, that is the number of days on which the subjects drank at all. For both men and women, the more often they drank the more they tended to drink on each day. Women tended to drink less frequently than men. For example, only 7.3% of women drank on all seven days, compared with 16.8% of men; whereas 25.5% of women drank on only one day of the week compared with 12.7% of men. Women also drank less on average per drinking day than men, at any given frequency of drinking (number of drinking days). For example, the geometric mean alcohol consumed by women who drank on all seven days of the week was 3.0 Units (24.0 gm = exponential of 3.18), whereas for men it was 5.7 Units (45.2 gm = exponential of 3.81). However the average amount drunk per day by women who drank more frequently was on a par with that of men who drank relatively infrequently. For example, women who drank on all seven days of the week drank about the same average amount per day (3.0 Units, 24.0 gm) as men who drank on only two days of the week (3.1 Units, 24.3 gm = exponential of 3.19).

Of the 128 (2^7) possible patterns of signs of drinking over the week, the most frequently occurring were those who did not drink at all (433, 21.6%), followed by those who drank on every day of the week (194, 9.7%). Of the other patterns those that involved drinking only on weekend days occurred most frequently. Altogether 379 (18.9%) subjects drank only at weekends. Few people drank only on weekdays (113, 5.6%), and these tend to drink infrequently: 88 of them drank on one day only and only one individual drank on all the weekdays. Amongst those who drank on both a weekend and a weekday, the only individual patterns which accounted for 20 individuals (1%) or more were those drinking on every day (194, 9.7%) or on six days of the week (174, 8.7%), or those drinking on the weekend and consecutive day(s) ('extended weekends'; 398 or 19.9%). In other words, people who drank both on a weekend day and a weekday tended to drink frequently.

Table 5.6: Mean alcohol consumption per drinking day (mean of logged average positive alcohol consumption, in gm, per drinking day) by drinking pattern (Completers only: $N = 2002$)

Pattern of drinking	*N*	%	Average alcohol consumption per drinking day
Weekday only	113	5.6	2.56
Weekend only	379	18.9	2.96
Weekday and weekend	1077	53.8	3.29
All who drank	1596	78.4	3.16
None	433	21.6	
All	2002	100.0	

Table 5.6 summarises the frequency of patterns according to whether people drank only on weekdays, only on weekends or on both a weekend and on a weekday. Those who drank only on weekends drank more heavily on average per drinking day than those who drank only on weekdays, but not as heavily as those who combined weekend and weekday drinking. People who drank on weekdays only tended to drink infrequently and moderately. Those who drank at weekends only drank relatively infrequently, but less moderately, whilst those who drank on both weekends and weekdays drank more frequently and more heavily.

5.6 The effect of the diary day order

The diary was not collected in the order of the days of the week, since it was started two days before the day the nurse interviewer happened to call on the subject. (The workload of the nurse was spread over all the days of the week, though she naturally tended to call during the week if possible, so fewer interviews were conducted at weekends.) Neither the total alcohol consumed during the diary week nor the missingness of the diary was associated with the day of the week on which the diary started (called the first day). (Results are not presented).

The importance of the diary day order is that different methods of data collection were used in the first two days (retrospectively completed by the nurse at the interview) and in the following five days (which the respondents were asked to complete for themselves at the time of eating or drinking). The relationship between the self-completed parts of the diary (last five days) and the part completed by the nurse (first two days) is shown in Table 5.7. It compares the percentage reporting any drinking and the amount of drink (mean of logged positive quantity) for days completed by the subject (self-completion) versus days completed by the nurse.

There is no consistent pattern in the differences of reporting alcohol consumption by self-completion versus completion by the nurse. More men reported drinking to the nurse than when they completed the diary themselves on every day of the week except Sunday (54.4% of men report drinking to the nurse on Sundays and 57.4% report drinking on Sundays when they complete this diary day themselves), but they tended to report heavier drinking to the nurse on that day (mean log of positive amount on Sunday is 3.55 to the nurse vs. 3.40 when they complete the diary day themselves). On the other days of the week the positive amounts reported by the men hardly differ between self-completion and completion by the nurse, except on Mondays they report on average lower amounts to the nurse (mean log of positive amount on Monday is 3.26 to the nurse vs. 3.38 when they complete the diary day themselves). On Thursdays, Fridays and Saturdays fewer women reported drinking to the nurse than they do when self-completing their diary records (e.g. on Friday 32.9% report drinking to the nurse whereas 39.1% do so when they self-complete that day), but slightly more do so on the other days of the week. For women, there are only small differences between the nurse completed mean positive amounts and those derived from self-completed records, and there is no consistent direction to the differences. There are in all 28 statistical tests (7 days × 2 measures × 2 for gender) and the P values produced look not unlike a random sample of P values. We would expect one or two P value less than 0.05 quite by chance. (A Bonferroni adjustment for the number of tests would indicate statistical significance only if $P < 0.05/28 = 0.002$, and none of the tests indicates such a small probability). In summary, the differences in reported alcohol consumption on days which were self-completed compared with those which were completed by the nurse are not consistent and they could have arisen simply by chance.

5.7 Discussion

Alcohol data in the diary is not missing completely at random. However we have a considerable amount of observed data in the NSHD 1989 survey that provide information about what people drank. A number of variables relate to the total alcohol consumption in the diet diary. The presence of this data allows us to model the alcohol consumption and to impute plausible values for the missing alcohol data provided we assume that the data are missing at random conditional on these variables (MAR). It would not be sensible to include all of these variables in an imputation model because they are mostly categorical and using them would subdivide the data into too many cells. Further, since they are inter-related, some cells will contain very few

Table 5.7: Comparisons of alcohol consumption self-completion versus completion by nurse

Percentage reporting any alcohol consumption self-completion versus completion by nurse

	Self Complete		Mon	Tue	Wed	Thu	Fri	Sat	Sun
Male	*Yes*	%	38.5	38.9	42.6	44.3	54.7	63.8	57.4
		N	678	664	727	758	810	828	726
Male	*No*	%	42.7	45.0	43.9	44.7	60.0	65.3	54.5
		N	527	585	497	421	325	323	462
		P	0.14	0.03	0.67	0.91	0.10	0.62	0.33
Female	*Yes*	%	22.8	24.2	27.5	27.1	39.1	49.9	40.3
		N	632	632	719	864	978	866	692
Female	*No*	%	23.1	24.5	31.5	26.4	32.9	44.8	41.8
		N	631	660	543	333	164	299	545
		P	0.88	0.89	0.13	0.82	0.14	0.13	0.59

Note: the *P*-values are obtained by χ^2 tests for independenceof proportions

Amount of alcohol consumed (mean log of positive quantity in gm) self-completion versus completion by nurse

	Self Complete		Mon	Tue	Wed	Thu	Fri	Sat	Sun
Male	*Yes*	Mean	3.38	3.30	3.27	3.35	3.53	3.69	3.40
		SD	0.85	0.89	0.85	0.91	0.88	0.89	0.95
		N	261	258	310	336	443	528	417
Male	*No*	Mean	3.26	3.31	3.28	3.32	3.56	3.69	3.55
		SD	0.89	0.85	0.89	0.95	0.84	0.90	0.97
		N	225	263	218	188	195	211	252
		P	0.12	0.83	0.93	0.68	0.69	0.90	0.04
Female	*Yes*	Mean	2.73	2.77	2.73	2.74	2.91	2.98	2.90
		SD	0.74	0.75	0.75	0.66	0.79	0.81	0.78
		N	144	153	198	234	382	432	279
Female	*No*	Mean	2.62	2.76	2.76	2.74	2.87	3.06	2.87
		SD	0.78	0.69	0.69	0.66	0.90	0.88	0.75
		N	146	162	171	88	54	134	228
		P	0.24	0.92	0.64	0.97	0.77	0.31	0.60

Note: the *P*-values are obtained by *t*-tests for differences of means

subjects, and they would provide insufficient information for estimating parameters. This issue will be discussed in Chapter 6. The most important variables to include are those that relate most directly to alcohol consumption and are also associated with non-response. These are: the weekly recall, the CAGE score, smoking status and current social class.

The object of the imputation is to complete the diaries that the subjects had failed to complete. This means imputing plausible values for alcohol consumption on each day of the week on which the diary was not completed. Since alcohol consumption had a distinct pattern depending on the day of the week, we need to find a way to take it into account. The pattern varied with gender and social class. Further, the pattern of signs was related to the pattern of amounts. The day of the week influenced not only whether people drank or not, but also how much they drank: people drank more on more popular drinking days. Individual patterns of signs over the week were associated with the amounts people drank. The more days someone drank the more they were likely to drink, and when they drank (at the weekends or during the week, or both) influenced the amount they drank. For most of the people who did not complete a diary we have data for the first two days, and if we can take into account the day of the week, this will provide the most direct information about their drinking.

The next chapter develops a method for dealing with item non-response to alcohol consumption in the diet diary that makes best use of the information examined here

Chapter 6

A Method for Dealing with Missing Data

6.1 Introduction

This chapter develops a method for dealing with item non-response that takes into account both the technical statistical problems which arise with such data and the characteristics of the data which are of substantive importance in alcohol research. Since the objective was to implement the method using existing software procedures (Section 1.4.2), developing the method involved critically evaluating the use of procedures for dealing with missing data which are available in standard software packages.

One of the problems in dealing with missing data is that we cannot verify how well a method performs in practice since we are by definition ignorant of the values of the items missing and, except where data is missing by design, of the process of non-response (the mechanism of missingness). This problem can be overcome by using a simulation in which we start with a complete dataset, delete some data values according to a known mechanism of missingness, and apply the method of dealing with the missing data to the incomplete dataset. The object of dealing with the missing data is not to recover the missing values themselves, but to make inferences which have good properties: that is, they are efficient and have small bias. To explore the properties of the methods, we examine the impact of applying them on estimates of interest — in this context, to alcohol research. Having applied a particular method, the estimates of interest are derived and can be compared with their known values derived from the complete data.

The theoretical deficiencies of different types of imputation method were discussed in the general taxonomy of methods (Section 2.7.2). These include considerations such as whether the method reflects the uncertainty about the missing values, preserves the associations among variables or makes full use of the data in the partially complete records. Nevertheless, such deficient methods are routinely used in practice, perhaps because they are available in standard software packages. This chapter explores how the methods work in practice for this application.

The simulations are used to explore the properties of the methods according to the following criteria:

1 How accurately they represent the measures of excessive alcohol consumption;

2 Whether they reflect the pattern of alcohol consumption over the days of the week

Since the aim is to develop a method for dealing with the diary data, similar realistically complex data was used in the simulation: the diary data for the 2002 respondents who

completed all seven days of their diet diary (completers).

Section 6.2 describes the way in which the simulations are conducted and how the results are assessed. The method of dealing with item non-response to alcohol consumption in the diet diary is developed in the subsequent sections, ordered according to the type of procedure used, as follows:

6.3 Naïve procedures: this section deals with the traditionally used procedures of ***Listwise Deletion***, and simple ***Mean Value Replacement***.

6.4 Procedures provided by the standard software package SPSS v 11.0 for single imputation: ***Regression*** and ***EM***.

6.5 Multiple imputation procedures provided by the specialist software package SOLAS™: ***Propensity Score***

6.6 Multiple imputation procedures provided by the specialist software package SOLAS™: ***Discriminant*** and ***Predictive Model Based Methods***.

6.7 Schafer's procedures ***CAT***, ***NORM*** and ***MIX***, provided by the software S-Plus as ***Loglin***, ***Gauss*** and ***Cgm*** (Conditional gaussian model), respectively.

Each section refers to the particular method being assessed, to which it applies the methods decribed in Section 6.2 below, and is structured so as to provide:

- A description of the method being assessed in that section;
- the results of the assessment;
- a discussion of the implications for the use of the method;
  and for the next stage in the development of the chosen method.

The sensitivity of some methods to the MAR assumption is assessed in Section 6.8.

Finally, a summary of the method which has been developed for application to the NSHD data is given in Section 6.9.

6.2 Methods

6.2.1 The simulation process

The 2002 respondents who completed the diet diary have all seven items of their alcohol data for the diary week observed and their diary record is referred to as the ***complete data***. An incomplete dataset is obtained from the complete data by randomly setting some data items to missing ('deletion') according to known mechanisms. Each method for dealing with missing data is applied to the incomplete data. The result is a dataset (or, for multiple imputation, m datasets) with no missing values either because cases with missing values have been ignored (i.e. the method uses complete cases only), or because the method has imputed values for the missing data. Such datasets will be referred to below as the ***completed data***.

Values of the proportions of subjects drinking over certain limits of alcohol consumption are calculated from the completed data, and compared with the same values calculated from the complete data.

Details of the process:

1 Take the complete data.

2 Randomly select data items to be set to 'missing' according to a particular mechanism of missingness (see Section 6.2.3, below).

3 Apply the method for dealing with missing data to the incomplete dataset from 2, to obtain completed data.

4 Calculate the values of the proportions drinking over certain limits from the completed data obtained in 3. (For multiple imputation, the values from the m completed datasets are first combined into a single value using the rules for combination given in Section 2.9).

5 Repeat **1–4** n times.

In this process we are not concerned with the original population from which the 2002 complete cases were sampled. We consider the original sampling process, and the process of missingness in the data, to be independent, so each contributes additively to the uncertainty, and here we are concerned with the uncertainty associated with the process of missingness. We assess this by estimating (as described below), from the n repetitions of the process defined in steps **1–5** above, the bias and the variability of the result of step **4** as an estimator of the value of the proportion calculated from the complete data.

Bias and Variability

Let p be a proportion calculated from the complete dataset of 2002 cases. Let

$$\hat{p}_1, \hat{p}_2, \cdots, \hat{p}_n$$

be the n values of the corresponding proportion calculated from the n repetitions of **1–4** above. For the random process defined in **1–5** above, each $\hat{p}$ is a random quantity with expectation $\mathrm{E}(\hat{p})$ and variance $V(\hat{p})$. The bias of $\hat{p}$ relative to p is

$$\mathrm{Bias}(\hat{p}) = \mathrm{E}(\hat{p}) - p$$

and its variance is

$$V(\hat{p}) = \mathrm{E}[(\hat{p} - \mathrm{E}[p])^2]$$

The mean square error of $\hat{p}$ is defined as

$$\begin{aligned}\mathrm{MSE}(\hat{p}) &= \mathrm{E}[(\hat{p} - p)^2] = \mathrm{E}[(\hat{p} - \mathrm{E}(\hat{p}) + \mathrm{E}(\hat{p}) - p)^2] \\ &= \mathrm{E}[\{\hat{p} - \mathrm{E}(\hat{p})\}^2] + (\mathrm{E}[(\hat{p}) - p)])^2 \\ &= V(\hat{p}) + \{\mathrm{Bias}(\hat{p})\}^2\end{aligned}$$

the root mean squre error of $\hat{p}$ is defined as

$$\text{RMSE}(\hat{p}) = \sqrt{\text{MSE}(\hat{p})}$$

and the standard error of $\hat{p}$ is given by

$$\text{SE}(\hat{p}) = \sqrt{V(\hat{p})} = \sqrt{\text{MSE} - \text{Bias}^2} = \sqrt{\text{RMSE}^2 - \text{Bias}^2}$$

From the results $\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_n$, these quantities can be estimated as

$$\text{Bias} = \frac{1}{n}\sum_{i=1}^{n}(\hat{p}_i - p) \tag{1}$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{p}_i - p)^2$$

$$\text{RMSE} = \sqrt{MSE} = \sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}(\hat{p}_i - p)^2\right)} \tag{2}$$

$$\text{SE} = \sqrt{\text{RMSE}^2 - \text{Bias}^2} \tag{3}$$

The hypothesis that the bias is zero is tested using the F statistic calculated from

$$F = \left(\frac{\text{Bias}}{\text{SE}}\right)^2 \tag{4}$$

with $(1, n-1)$ degrees of freedom (where n is the number of replications of the process). Confidence intervals (95%) are calculated using the t distribution with $(n-1)$ degrees of freedom.

The bias and its standard error depend on two aspects of the process defined in steps **1–5** above:

1. The missingness mechanism (MCAR, MAR, MNAR) that was used for deletion, to generate the 'missing' data
2. The method used for dealing with the missing data, which has two components:
   i. the procedure, and
   ii. the way in which the procedure is applied

6.2.2 Assessing the methods

The less the bias the better the method. For simplicity, a method that gives rise to bias in the process above is called a biased method. First, bias is assessed under MCAR (using the data which results from MCAR deletion). Any method should be unbiased under MCAR. If the method is not unbiased under MCAR the reason is further investigated using simple simulated data. The bias may arise because of a problem with the procedure itself (2 i. above) or because of problems posed by the suitability of the procedure for the data in this application (2 ii. above), e.g. the robustness to the semicontinuous distribution of alcohol consumption.

For imputation methods which are unbiased under MCAR and MAR, a further consideration is whether the method is able to reflect the pattern of alcohol consumption over the days of the week. This is not an issue with the procedure itself but with adapting the way the procedure is used to take into account the day of the week. This is assessed by comparing patterns of sign and amount of alcohol consumption in the data imputed under MCAR with the patterns in the complete data.

None of the methods is intended to work under MNAR. However, in reality we cannot know that data is not MNAR. It is claimed that, given a rich set of covariates, multiple imputation has the potential to protect against the impact of MNAR (Section 1.3.4). The results under MNAR are used to assess the sensitivity to the MAR assumption.

The details of the assessment method are given below.

6.2.3 Simulated mechanisms of missingness

The deletion is done so as to give similar proportions of missing items to those found in the original data set of diary data (for the 3262 respondents to the survey at age 43), stratified by gender (for MAR and MNAR). Most of the 3262 respondents recorded either 0, 2 or 7 days of their diaries, very few recording 1, 4, 5 or 6 days. (Section 3.2.3, Table 3.1, Figure 3.1). For simplicity either all seven days, diary days 3–7, or no days are deleted in the following proportions. For men: approximately 4.2% had all their seven diary day records deleted, 36% had days 3–7 deleted and 59.8% had none of the seven diary day records deleted. For women: approximately 2.6% had all seven days deleted, 34.4% had days 3–7 deleted, and 62.9% had no days deleted.

The simulated mechanisms of missingness are implemented as follows:

MCAR

Records are deleted using random selection of completers.

MAR

Observed records are deleted with probability proportional to the logarithm of reported alcohol consumption in the previous week (*weekly recall*). Of those (40 cases) with no recorded weekly recall, a random sample of seven cases have all diary records deleted, and eleven have days 3–7 deleted; the remaining 22 have no records deleted.

MNAR

The observed records are deleted with probability proportional to the reported alcohol consumption on the heaviest drinking day in the days to be deleted. For those who had all days deleted the maximum consumption in days 1 and 2 is used; for those who had days 3–7 deleted the maximum consumption in days 3–7 was used.

Details of Methods for MCAR, MAR and MNAR

Set 1 refers to the cases which are selected to have all diary records deleted. ***Set 2*** refers to the cases which are selected to have days 3–7 deleted, ***Set 3*** refers to the cases which are selected to have no days deleted.

MCAR

The proportion of records deleted does not depend on gender. First generate 2002 random numbers (r) uniformly distributed on (0,1). The cases corresponding to $r \leq 0.031$ go into Set 1. The cases corresponding to $0.031 < r \leq 0.390$ go into Set 2. The rest go into Set 3.

MAR* and *MNAR

Generate random numbers numbers (r) uniformly distributed on (0,1) separately for men and for women.

Set $p = W/C$, where

W = log(1 + previous week's drink total in gm) for MAR

W = log(1 + maximum consumption over relevant days) for MNAR

C = a constant, varying for each Set 1 and 2, found empirically as below

Then select cases for which $r \leq p$.

C is found by trial and error so that the expected number of cases (the mean in a 1000 runs) is approximately that required for the particular set being selected, according to the proportions in the first paragraph of this section.

First, Set 1 is selected. Then Set 2 is selected, by selecting appropriate numbers for Set 1 + Set 2, and discarding any cases that are already selected for Set 1. The remainder go into Set 3.

Then the diary records for all days 1–7 are deleted for cases in Set 1, records for days 3–7 are deleted for cases in Set 2, and none of the seven daily records are deleted for the cases which remain (Set 3).

6.2.4 Measures of alcohol consumption

The measures of alcohol consumption of interest for alcohol research are the proportions of respondents, by gender, who drink over the recommended limits, defined as follows:

1. Weekly limits (Royal College of Psychiatrists, 1986):
   A. Women drinking over 14 units per week, men over 21 units per week. This level of alcohol consumption is referred to as ***excessive***.
   B. Women drinking over 35 units per week, men over 50 units per week. This level of alcohol consumption is referred to as ***heavy***.

and

2. Daily limits (Faculty of Public Health Medicine, 1996):
   A. Women drinking over 3 units in a day, men over 4 units in a day, on at least one day of the week. This level of alcohol consumption is referred to as ***excessive***.
   B. Women drinking over 6 units in a day, men over 8 units in a day, on at least one day of the week. This level of alcohol consumption is referred to as ***heavy***.

Note that excessive drinking includes both immoderate and heavy drinking.

6.2.5 Modelling alcohol consumption

Except for naïve procedures, alcohol consumption (in grams) on each of the seven diary days (Section 3.2) is modelled using a set of covariates. The choice of covariates in these models is based on the conclusions of earlier analysis. In Section 5.7 it was concluded that the most important variables to include were the weekly recall (in Units) (Section 3.2), the CAGE score (Section 3.2, classified as 0, 1 or 2–4), smoking status (smoker or non-smoker as described in Section 4.2) and adult social class (classified as manual or non-manual, as described in Section 5.2), the day of the week, and the alcohol consumption on the diary days completed.

6.3 Naïve methods

6.3.1 Introduction

This section evaluates the methods of **Listwise Deletion** (LD) based on the complete records, and **Mean Value Replacement** (MVR); see Section 2.8.2.

6.3.2 Methods

Complete cases only or listwise deletion (LD)

Only the cases with complete seven daily records are used. All the incomplete records are discarded and not used in the analysis at all.

Mean value replacement (MVR)

Missing data is replaced by a single imputed value using the mean value for the gender group on the diary day.

These methods were implemented in SPSS v11.0. LD and MVR were each applied to 100 repetitions of the simulation process ($n = 100$, Section 6.2.1) using the MCAR deletion mechanism (Section 6.2.3). LD was also applied to 100 repetitions using the MAR and MNAR deletion mechanisms (Section 6.2.3).

6.3.3 Results

The estimates of the percentages drinking over the chosen weekly and daily limits (Section 6.2.4) are first derived from the complete data (before any data are set to missing), consisting of the 2002 respondents who completed their diary. These are given in the first column of Table 6.1. For example, 15.6% of women drank in excess of 14 Units of alcohol in total during the week, while 38.1% of them drank more than 3 Units on any day during the week. In the context

of the simulation of the non-response process, these complete data proportions are considered, for reference purposes, to be fixed.

We delete some items at random, using the MCAR mechanism, so there are now missing values in the dataset. Let us take, for example, the method MVR and the estimated proportion of women drinking over 14 Units in the week. We apply MVR to complete the dataset and calculate the proportion of women drinking over 14 Units in the week from this completed dataset. This process is repeated 100 times, each time using a different set of MCAR data. On average, according to the datasets completed using MVR, the proportion of women drinking over 14 Units of alcohol is only 10.75%. So on average, the bias in this estimate compared to the known value of 15.6% is −4.85 (Table 6.1). More technically, the bias is estimated as the mean of the differences between the estimates derived from the method and that from the complete data, as given in equation (1) in Section 6.2.1. This negative bias (−4.85) seems substantial. However, we need to estimate the error in this estimate of bias. The standard error is only meaningful as a measure of the accuracy of estimation for unbiased estimators. So we use the root mean square error (RMSE) to measure the overall error, including bias and variability (calculated by equation (2) in Section 6.2.1). The standard error (SE) is calculated from the bias and RMSE (by equation (3) in Section 6.2.1). The standard error of our estimated bias is 0.74, with 95% confidence interval −6.32 to −3.38. Hence we can be 95% confident that using MVR gives a negative bias in the estimated proportion of women drinking over 14 Units of alcohol during the diary week. The F-test can be used to test the hypothesis that the bias is zero. The F-statistic (calculated by equation (4) in Section 6.2.1) is significant at the 0.1% level.

The results in Table 6.1 show that LD gives unbiased estimates when the data is MCAR. MVR gives substantially and significantly downwardly biased estimates even under MCAR. This is because for men and for women the limits considered were higher than their mean alcohol consumption, so when missing values are replaced by mean values these are all below the limits, resulting in underestimation of proportions exceeding the limits.

The LD estimates are substantially and significantly negatively biased under mechanisms that depart from MCAR, as shown in Table 6.2. For example, the proportion of women drinking over 3 Units on any day of the week is estimated, on average, as 38.3% under MCAR (38.1 + 0.22), only 25.9% under MAR (38.1 − 12.24), and 23.1% under MNAR (38.1 − 15.01). Using the MAR and MNAR mechanisms defined in Section 6.2.3, heavier drinkers are more likely to be deleted. Using the MAR mechanism, those with heavier weekly recalled alcohol consumption are more likely to have their diary records deleted, and these respondents tend to drink more during the diary week (Section 5.3). Using the MNAR mechanism, deletion of diary day records is directly related to the level of drinking on the days deleted. Hence the LD method, which ignores the people with missing data, results in large negative bias under MAR and even greater negative bias under MNAR. These results confirm that the simulated mechanisms of missingness were implemented correctly.

Table 6.1: Estimates by listwise deletion and mean-value replacement on MCAR data of proportions drinking over weekly and daily limits: comparison with estimates from complete data.

	Complete	LD					MVR				
	data (%)	BIAS	RMSE	SE	95% CI	*F*	BIAS	RMSE	SE	95% CI	*F*
Estimated proportions with alcohol consumption over weekly limits											
Women											
>14 U	15.6	0.07	0.86	0.85	(−1.62,1.76)	0.01	−4.85	4.91	0.74	(−6.32, −3.38)	42.8
>35 U	1.1	0.01	0.27	0.27	(−0.53,0.55)	0.00	−0.43	0.46	0.15	(−0.73, −0.13)	8.2
Men											
>21 U	33.2	0.04	1.25	1.25	(−2.44,2.52)	0.00	−3.77	3.91	1.02	(−5.79, −1.75)	13.8
>50 U	8.6	−0.10	0.71	0.70	(−1.49,1.29)	0.02	−3.03	3.06	0.42	(−3.86, −2.20)	51.8
Estimated proportions with alcohol consumption over daily limits											
Women											
>3 U	38.1	0.22	1.16	1.14	(−2.04,2.48)	0.04	−9.06	9.09	0.76	(−10.57, −7.55)	141.4
>6 U	11.5	0.15	0.77	0.76	(−1.36,1.66)	0.04	−2.87	2.91	0.47	(−3.80, −1.94)	38.0
Men											
>4 U	63.3	−0.04	1.10	1.10	(−2.22,2.14)	0.00	−11.94	11.97	0.88	(−13.69,−10.19)	182.8
>8 U	35.0	−0.11	1.19	1.18	(−2.45,2.23)	0.01	−7.90	7.93	0.62	(−9.13, −6.67)	161.2

Percentage points of $F_{1,99}$: $P = 0.05$ $F = 3.94$, $P = 0.01$ $F = 6.90$, $P = 0.001$ $F = 11.5$

Note: In the tables in this chapter, levels of alcohol consumption are indicated as follows:

Exceeding weekly limits:

Excessive drinking

Women	**Men**	
>14 U		More than 14 units consumed in total during the week
	>21 U	More than 21 units consumed in total during the week

Heavy drinking

Women	**Men**	
>35 U		More than 35 units consumed in total during the week
	>50 U	More than 50 units consumed in total during the week

Exceeding daily limits:

Excessive drinking

Women	**Men**	
>3 U		More than 3 units consumed in a day, on at least 1 day
	>4 U	More than 4 units consumed in a day, on at least 1 day

Heavy drinking

Women	**Men**	
>6 U		More than 6 units consumed in a day, on at least 1 day
	>8 U	More than 8 units consumed in a day, on at least 1 day

NB: 'Excessive drinking' includes 'Heavy drinking'

6.3.4 Discussion

Most statistical procedures in widely used software packages apply listwise deletion by default. Since this method of dealing with missing values is automatically implemented by the software, it requires no effort on the part of the user, who may not even be aware of the fact that the analysis is based only on the subset of cases with complete information for the variables in the analysis. In some software packages other naïve options, such as mean value replacement (MVR), are available and can be implemented without effort. For example, MVR is available in the SPSS Linear Regression or Factor Analysis procedures simply by selecting it as an option in the dialogue box.

Listwise deletion yields unbiased estimates when the data is missing completely at random (MCAR) since the cases remaining after MCAR deletion are a random sample of the original cases. However even under MCAR this method is inefficient since it is based on a smaller sample. If our data is MAR or MNAR the consequence of ignoring missing data is that estimates may be seriously biased.

Mean value replacement (MVR) gives downwardly biased estimates even under MCAR. There are instances in which MVR would not produce biased estimates, for example if used to estimate the mean value of a Normally distributed variable. Replacement with the median value is sometimes used when the distribution is skewed. In the case of alcohol consumption even the median may be at the lowest level of drinking measured (see, for example, Gmel, 2001). The gender specific median alcohol consumption is below the weekly limits of interest and below all except the lower daily limit for men (see Table 6.1, Complete data (%)). So median replacement is not a suitable method to use here.

However, even if the variable with missing values is Normally distributed, MVR (and also median value replacement) has additional serious theoretical disadvantages (see Section 2.7.2). By using MVR the variance of a variable is under-represented because the imputed values do not deviate from the mean. In the process, the correlations with other variables used in the analysis are also distorted, because the imputation is unaffected by the values of these variables. Moreover, MVR replaces each missing value with just one imputation (single imputation) that is treated in the completed data set as though it were observed. The standard error of any estimate based on the (single) completed data set is underestimated because it does not take into account the uncertainty due to missingness and pretends that more data is available than was in fact collected.

Table 6.2: Estimates of proportions over weekly and daily limits from listwise deletion on MCAR, MAR and MNAR data: comparison with proportions estimated from complete data

Estimated proportions with alcohol consumption over weekly limits

	Complete data	MCAR					MAR					MNAR				
Women	%	BIAS	RMSE	SE	95% CI	*F*	BIAS	RMSE	SE	95% CI	*F*	BIAS	RMSE	SE	95% CI	*F*
>14	15.6	0.07	0.86	0.85	(−1.62,1.76)	0.01	−7.49	7.54	0.82	(−9.12,−5.86)	84.2	−7.88	7.92	0.74	(−9.35, −6.41)	113.3
>35	1.1	0.01	0.27	0.27	(−0.53,0.55)	0.00	−0.68	0.71	0.21	(−1.10,−0.26)	10.5	−0.77	0.80	0.23	(−1.23, −0.31)	11.6
Men																
>21	33.2	0.04	1.25	1.25	(−2.44,2.52)	0.00	−8.01	8.08	1.03	(−10.05,−5.97)	60.4	−9.93	10.03	1.38	(−12.67, −7.19)	51.9
>50	8.6	−0.10	0.71	0.70	(−1.49,1.29)	0.02	−2.86	2.94	0.69	(−4.23,−1.49)	17.4	−3.35	3.42	0.69	(−4.72, −1.98)	23.8

Estimated proportions with alcohol consumption over daily limits

	Complete data	MCAR					MAR					MNAR				
Women	%	BIAS	RMSE	SE	95% CI	*F*	BIAS	RMSE	SE	95% CI	*F*	BIAS	RMSE	SE	95% CI	*F*
>3	38.1	0.22	1.16	1.14	(−2.04,2.48)	0.04	−12.24	12.3	1.17	(−14.56,−9.92)	108.9	−15.01	15.06	1.16	(−17.31,−12.71)	166.7
>6	11.5	0.15	0.77	0.76	(−1.36,1.66)	0.04	−5.21	5.25	0.66	(−6.52,−3.90)	61.5	−5.61	5.65	0.61	(−6.82, −4.40)	84.3
Men																
>4	63.3	−0.04	1.10	1.10	(−2.22,2.14)	0.00	−8.91	8.95	0.81	(−10.52,−7.30)	122.5	−12.92	13.00	1.44	(−15.78,−10.06)	80.0
>8	35.0	−0.11	1.19	1.18	(−2.45,2.23)	0.01	−7.16	7.25	1.11	(−9.36,−4.96)	41.5	−10.36	10.44	1.24	(−12.82, −7.90)	70.0

Percentage points of $F_{1,99}$: $P = 0.05$ $F = 3.94$, $P = 0.01$ $F = 6.90$, $P = 0.001$ $F = 11.5$

6.4 Methods using SPSS procedures

6.4.1 Introduction

This section evaluates the use of the procedures available in SPSS software (Section 2.8.3): Regression (Section 2.8.3.1) and EM (Section 2.8.3.2).

SPSS, widely used by epidemiologists, now offers two procedures for imputation, which enable the analyst to exploit the information in incomplete records (see Section 2.7.2). The procedures in SPSS impute for missing values of a variable with missing data (Y) by modelling Y based on the observed values of the variables to be imputed and a set of covariates (X), assuming that Y is missing at random (MAR) conditional on these covariates.

6.4.2 Methods

The variables used in the model for these imputation procedures are given in Section 6.2.5. The procedures for imputation in SPSS are designed for imputing continuous variables only and not for categorical covariates (Section 2.8.3). The covariate CAGE score (classified as 0, 1 or 2–4) is first coded as two binary dummy variables, and the day of the week of the first diary day is coded as 3 binary dummy variables. The Regression method assumes that the variables have a multivariate Normal distribution. This is the default assumption for SPSS EM also (and is used here). Regression imputes one diary item at a time using the other diary items, and all the covariates, as independent variables. SPSS EM imputes all the variables with missing values at the same time.

These methods were implemented in SPSS v11.0. Regression and EM were each applied to 100 repetitions of the simulation process ($n = 100$, Section 6.2.1) using the MCAR deletion mechanism (Section 6.2.3). In the application of the procedures to this data some imputations for alcohol consumption are negative. Any negative imputed values of alcohol consumption are set to zero before calculating the estimated totals. Both procedures produce only single imputations.

6.4.3 Results

Table 6.3 gives the estimates from the SPSS procedures Regression and EM on MCAR data. Although the magnitude of the biases for both these methods is lower than the corresponding biases for MVR (Table 6.1), they each have substantial biases for some of the estimates. Biases are relatively greater for estimates over daily limits than for those over weekly limits. For estimates over daily limits, SPSS EM produces substantially negatively biased results, which are statistically significant at the 0.1% level, whereas SPSS Regression produces large (and significant) positively biased results for women and small (and non-significant) biases for men. It is interesting that the biases that are substantial are in opposite directions for the two procedures.

Table 6.3: Estimates from SPSS Regression Method and SPSS EM on MCAR data.
Comparison with estimated proportions over weekly and daily limits from complete data

Complete		SPSS Regression					SPSS EM				
data (%)		BIAS	RMSE	SE	95% CI	*F*	BIAS	RMSE	SE	95% CI	*F*
Estimated proportions with alcohol consumption over weekly limits											
Women											
>14	**15.6**	2.22	2.34	0.76	(0.71, 3.73)	8.5	−0.66	0.84	0.52	(−1.69, 0.37)	1.6
>35	**1.1**	0.27	0.36	0.24	(−0.21, 0.75)	1.3	0.07	0.21	0.20	(−0.33, 0.47)	0.1
Men											
>21	**33.2**	0.67	1.06	0.81	(−0.94, 2.28)	0.7	−0.73	0.91	0.53	(−1.78, 0.32)	1.9
>50	**8.6**	−0.24	0.50	0.44	(−1.11, 0.63)	0.3	−0.45	0.58	0.36	(−1.16, 0.26)	1.5
Estimated proportions with alcohol consumption over daily limits											
Women											
>3	**38.1**	7.73	7.81	1.11	(5.53, 9.93)	48.4	−7.52	7.56	0.70	(−8.91, −6.13)	116.9
>6	**11.5**	5.01	5.10	0.97	(3.09, 6.93)	26.6	−2.84	2.86	0.39	(−3.61, −2.07)	51.9
Men											
>4	**63.3**	1.35	1.93	1.37	(−1.37, 4.07)	1.0	−8.41	8.44	0.72	(−9.84, −6.98)	137.6
>8	**35.0**	−1.44	1.80	1.09	(−3.60, 0.72)	0.7	−7.07	7.10	0.64	(−8.34, −5.80)	123.0

Percentage points of $F_{1,99}$: $P = 0.05$ $F = 3.94$, $P = 0.01$ $F = 6.90$, $P = 0.001$ $F = 11.5$

Further diagnostic tests were conducted to investigate the reason for the bias: whether the bias resulted from non-normality in the distribution of alcohol consumption, or from features of the algorithms used by the procedures.

6.4.4 Further diagnostic tests

Introduction

In order to determine whether the problems with the methods resulted from the semicontinuous distribution of alcohol consumption (see Section 2.4), the procedures were tested using two simply constructed simulated datasets. The first dataset consisted of Normally distributed variables, the second consists of semicontinuous variables consisting of two parts: zeros and log-Normal, in proportions similar to those in diaries of alcohol consumption.

There are three correlated variables in each dataset and some values of each variable are deleted completely at random, so that only one variable is missing for each case. The missing data is then imputed using the studied procedure (Regression or EM).

Method

Each simulated dataset consists of 300 cases of three correlated variables, X_1, X_2, X_3 (the complete data). The details of how these are constructed are given below. These correlated variables are copied to variables named Y_1, Y_2 and Y_3 (respectively) from which 100 values of each variable are deleted MCAR so that only one variable is missing for each case. The set of simulated variables with missing data, Y_1, Y_2 and Y_3, has known joint distribution. The data presents a simple task for the imputation procedure since for each deleted value there are observed values in two correlated covariates. Y_1, Y_2 and Y_3 are used in a model to impute the missing values using the studied procedure. The resulting completed data Y_1, Y_2 and Y_3, correspond to the complete data in X_1, X_2 and X_3, respectively. Hence Y_1 consists of two sets of values: the observed values, which are the same as those in X_1 where these had not been deleted (200 values), and the imputed values, which differ (in general) from the corresponding values of X_1 which had been deleted (100 values). Similarly for Y_2 and Y_3.

For each dataset and procedure, the results are assessed by plotting a graph of the X variable against the corresponding Y variable, for example X_2 and Y_2. For the observed values of Y_2 the points will lie on a straight line ($Y_2 = X_2$). For the imputed values of Y_2 the points will (in general) be scattered around this line (since then $Y_2 \neq X_2$). Thus, in the graph, the imputed values are easily distinguished from the observed values. If the imputation procedure has succeeded in preserving the sampling variability of Y (Y_2 in this case) then the plot of the imputed Y against the original X should exhibit an appropriate bivariate dispersion about the straight line ($Y_2 \neq X_2$).

Construction of the simulated datasets

Normally distributed variables: The first simulated dataset consists of 300 cases of 3 Normally distributed correlated variables, X_1, X_2 and X_3, constructed using the following procedure:

1. Start with independent identically distributed Normal variables U_1, U_2 and U_3:
   U_1 is a random sample of 300 from a Normal distribution, mean = 0, standard deviation = 1.
   U_2, U_3 are similar random samples.
2. Next set $X_1 = U_2 + U_3$, $X_2 = U_1 + U_3$, $X_3 = U_1 + U_2$.
   X_1, X_2 and X_3 each have mean 0, variance 2. The covariance of each pair of variables X_1, X_2 and X_3 is 1, and their correlation is 0.5.
3. Next, $X_1 = X_1 + 1$, $X_2 = X_2 + 2$, $X_3 = X_3 + 3$,
   so that the means of X_1, X_2, X_3 are 1, 2, 3 respectively.
4. Finally, Y_1, Y_2, Y_3 are the same as X_1, X_2, X_3 but cases 1–100 of Y_1, 101–200 of Y_2 and 201–300 of Y_3 are set missing ("NA").

Semicontinuous distributed variables: The second set of simulated data was generated in a similar way to the first set except that the variables X_1, X_2, X_3, had a semicontinuous

distribution, with approximately 60% of cases in each X being set to zero at random (similar to the proportions in the NHSD on any diary day) and the remainder are log Normal (i.e. the values of X are transformed by the exponential function).

Results

The results from the Normal data are illustrated in the graphs of X_2 against Y_2 given in Figure 6.1. Each graph shows the plot of the complete data for X_2 against the completed observed and imputed values of this variable (Y_2). Figure 6.1 Graph 1 illustrates the results using SPSS Regression, in which the imputed values show an appropriate scatter around the line. Figure 6.1, Graph 2 shows the results using SPSS EM algorithm in which the imputed values are not scattered randomly about the line. The imputations are consistently greater than X for low values of X and lower than X for high values of X. (These results exhibit regression towards the mean since the ideal regression of the imputed value on the observed value which it had before deletion theoretically has slope 1/3). The output for EM (Figure 6.1, Graph 2) can be compared with that from Schafer's procedure NORM, described in Section 2.8.5, which also uses the EM algorithm. The output after imputation using Schafer's NORM (Figure 6.1, Graph 3) shows appropriate scatter of the imputed values, similar to that for SPSS Regression. A further test using SPSS Regression method with no random component added for the imputations (Figure 6.1, Graph 4) shows a similar result to that of SPSS EM.

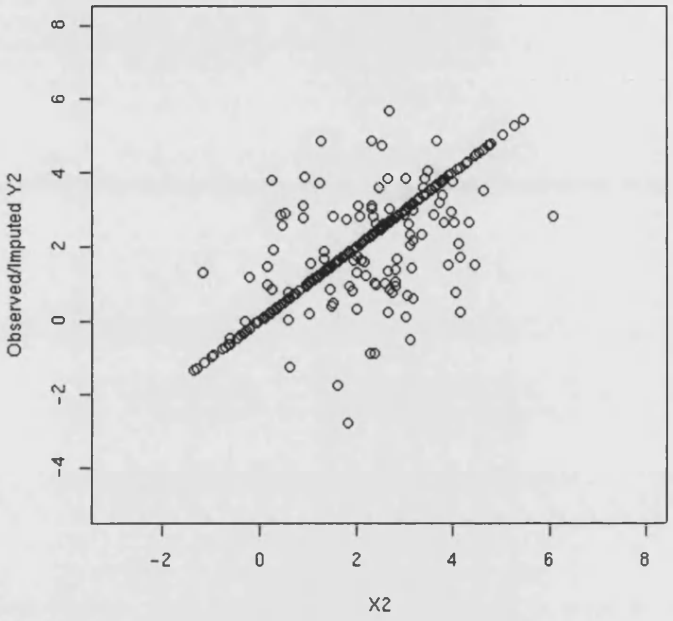
SPSS Regression produces satisfactory results with Normally distributed data but it is necessary to see how it copes with the semicontinuous distribution characteristic of alcohol data. The output from SPSS Regression using the semicontinuous data is shown in Figure 6.2, Graph 1. This shows that the method imputes a wide scatter of values (Y_2) where the true value (X_2) is zero and in addition there is a tendency for high values of imputations (Y_2) for low positive values of X_2, and low values of imputations (Y_2) for high values of X_2. Similar results are produced by Schafer's NORM applied to the same data (Figure 6.2, Graph 2). The results (not shown) are similar if all the (semicontinuous) X_2 values, including the zeros, are transformed using the logarithm of (1+ X_2), indicating that using such a transformation would not resolve the problem of the semicontinuous distribution of the alcohol data.

[Text continues following Figures 6.1 and 6.2 below]

Figure 6.1 Graphs of complete data values for X_2 against observed and imputed values of this variable (Y_2) using simulated normally distributed data.

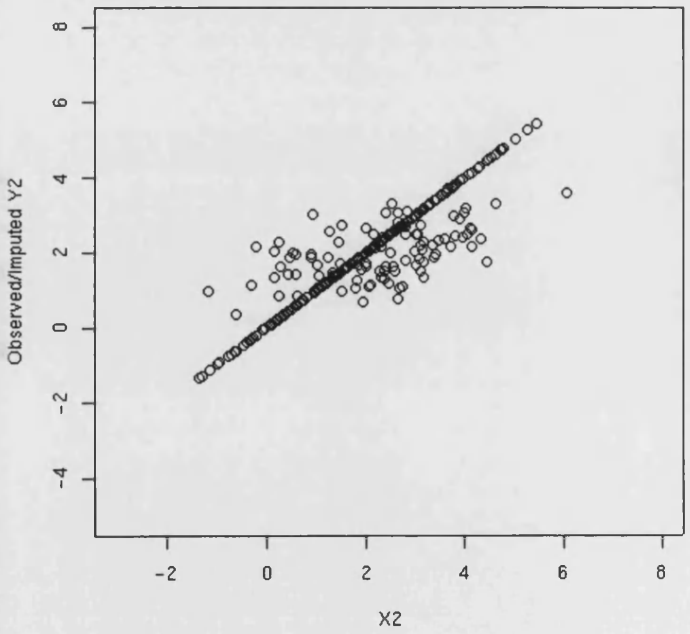
Graph 1

SPSS Regression with random error term, (random normal deviate, scaled by standard error of estimate)



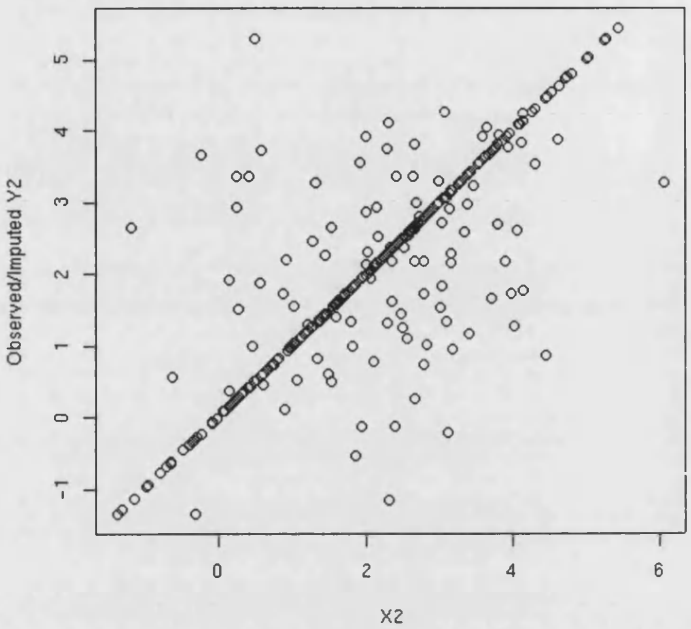
Graph 2

SPSS EM (no random component can be added)



Graph 3

Schafer's NORM (using EM algorithm)



Graph 4

SPSS Regression with no random component added

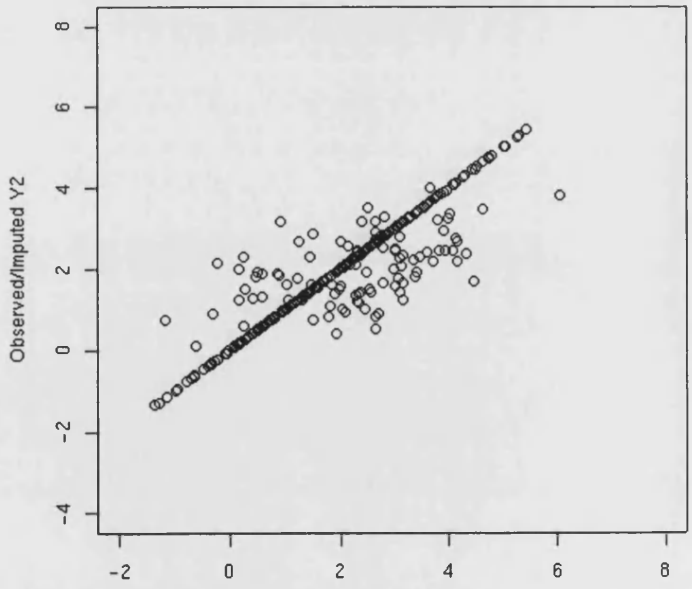
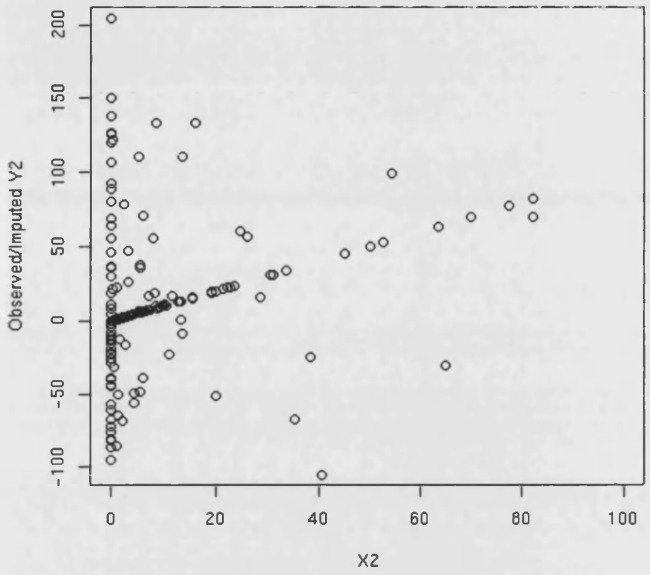


Figure 6.2 Graphs of complete data values for X_2 against observed and imputed values of this variable (Y_2) using simulated semicontinuous data.

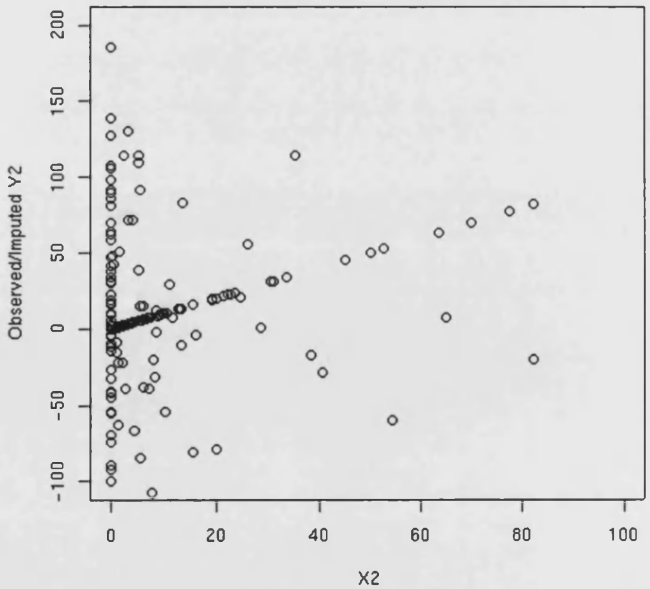
Graph 1

SPSS Regression with random error term, (random normal deviate, scaled by standard error of estimate)



Graph 2

Shaffer's NORM (using EM algorithm)



Conclusions

SPSS Regression produces a satisfactory spread of imputations with Normally distributed data, whereas SPSS EM does not. It was explained in Section 2.8.3.2 that the SPSS EM procedure does not allow a random component to be added to the imputations. The comparison of the results from SPSS EM with those of Schafer's NORM and Regression without a random component confirms that the inappropriate distribution of the imputed variable is explained by the lack of a random component in SPSS EM. In this procedure, the imputed values are taken to be the expected values in the final E step. SPSS EM underestimates the variance of the imputed variable (here Y_2). These results explain the negative bias in the estimates of proportions over limits using SPSS EM shown in Table 6.3.

SPSS Regression (and Schafer's NORM) does not produce an appropriate distribution for the imputations when the variable to be imputed is semicontinuous. Nor does a logarithmic transformation including the zeros solve this problem, probably because of the high proportion of zeros in the data. When the true value (X_2) is zero, imputation of large positive values of Y_2 will give an increase in the proportion of high imputed values for Y_2 relative to X_2. This explains the positive bias that was obtained using SPSS Regression shown in Table 6.3.

6.4.5 Discussion

Neither of the SPSS methods for imputation of missing values, Regression nor EM, is suitable for the imputation of alcohol consumption. Several of the estimates of excessive drinking are substantially and significantly biased when the data is MCAR (Table 6.3). By applying the methods to very simple simulated data, their shortcomings have been identified in Section 6.4.3.

Both methods failed to produce appropriate imputations for semicontinuous distributed data which is characteristic of alcohol consumption, but Schafer's NORM also has this drawback. This failure may be expected since all these procedures assume that the data has a multivariate Normal distribution. The SPSS Regression procedure (and Schafer's NORM) performs appropriately with Normally distributed data, so it could be used in applications where this assumption is reasonable. SPSS EM underestimates the variance in the variable to be imputed and is not appropriate even for Normally distributed data.

Another serious shortcoming of both methods is that they produce only single imputations and so they do not take into account the uncertainty due to missing values (Section 2.7.2). The Regression method could be adapted to generate several imputations by rerunning the procedure m times to produce m completed datasets. This is possible because SPSS Regression gives a stochastic imputation (Section 2.7.2). The Regression procedure has provision for adding a random component to the imputations, and if this option is used the imputations are not unique. SPSS EM, on the other hand, has fundamental methodological problem: it is deterministic. In the SPSS EM procedure, the imputed values are taken to be the expected values of the data obtained in the final E step (see Section 2.8.3.2). This is not a characteristic of the EM algorithm itself but of its implementation by SPSS. (Schafer's NORM, which uses the EM

algorithm, produces stochastic imputation, and preserves the sampling variability in the variable with missing values). It is not clear why there is no provision for adding a random component to the imputations in EM (communications with SPSS support and with SPSS email user group failed to produce any answers), and its absence accounts for the failure of the SPSS EM to preserve the sampling variability in the variable with missing values, even with Normally distributed data.

The approach used in SPSS Regression has a disadvantage compared with using the EM algorithm properly as a basis for imputation, as it is in Schafer's NORM. SPSS Regression uses all the observed data in the partially complete records, but not at the same time. It imputes one variable with missing values (diary item) at a time using pairwise available values. It cannot make full use of the data in the partially complete records, nor take full account of the associations between the variables used.

6.5 SOLAS Propensity Score

6.5.1 Introduction

One way to avoid the problem posed by the semicontinuous distribution of alcohol consumption is to use a method of imputation which does not make assumptions about the distribution of the variable to be imputed. One such method is the Propensity Score, implemented in the specialist software for missing data, SOLAS. Propensity Score uses Logistic Regression to model the missingness of the variable (a binary indicator) rather than the value of the variable itself, and so is non-parametric (Section 2.7.2). This section evaluates the use of the Propensity Score procedure in SOLAS software. This is a procedure for multiple imputation. The general approach of SOLAS is explained in Section 2.8.2 and the details of the Propensity Score procedure are given in Section 2.8.4.1.

6.5.2 Methods

Briefly, the propensity score is the conditional probability of missingness given the observed covariates (Rosenbaum and Rubin, 1983). In SOLAS, a propensity score is generated separately for each variable with missing values by fitting a logistic regression model using observed covariates. The imputations are randomly drawn from a sample of the observed responses for cases that have similar propensity scores (called the donor pool). In this application, the donor pool was defined by quintiles of the propensity score. The user can chose other options for the donor pool and these are given in Section 2.8.4.1 procedure step 5.

The variables used are those specified in Section 6.3.2. The software automatically generates design variables for any nominal categorical covariates. In SOLAS, missing values in covariates must first be imputed (Section 2.8.4). The missing values in weekly recall and CAGE are imputed by hot deck conditioning on gender and smoking status.

The process above is repeated to give m independent samples, producing m imputations for each missing value and m completed datasets (m is chosen by the user). The proportions of interest are calculated from each of m completed datasets and are combined to give the MI-estimate (equation (1) in Section 2.9) and the MI-estimate of the variance (equation (4) in Section 2.9). In this application five imputations are used ($m = 5$).

The procedure was applied to only one set of simulated MCAR data. The procedure was originally implemented using SOLAS V 1.0, during the early part of the investigation. The software was re-released, currently as version 3.2, but using this release to fit the same model the programme crashed. The reason for this crash was not investigated further, since the method was unsuitable for use in this application, and the results from one run are used simply to illustrate its properties. Since the process was not repeated, the bias and its standard error cannot be estimated.

6.5.3 Results

The procedure uses Multiple Imputation (MI), and the process by which the MI-estimate and its standard error are derived is first explained.

MI-estimates

Five imputations are made for each missing value. One imputation for each of the missing data values is used to make up a completed dataset, giving five different completed datasets. These completed datasets are analysed using complete-data methods. In this instance, we simply derive the proportions of interest from each of the five datasets. To illustrate the method we use the example of women drinking excessively during the week. The proportion and its variance from the five completed datasets, and their MI-estimates are given in Table 6.4.

Table 6.4: Percentage of women (n=1024) consuming more than 14 units of alcohol in the diary week from 5 completed datasets (Propensity score MI)

	Completed dataset					
	1	2	3	4	5	MI-estimate
%	19.24	18.65	20.61	18.95	19.14	19.32
Variance (%)	1.52	1.48	1.60	1.50	1.51	2.20

The MI-estimate (equation (1) in Section 2.9) is the mean of the percentages from each completed dataset, in this example 19.32. It can be seen that the MI-estimate of the variance is greater than the variances for each dataset. This is because it takes into account the data incompleteness. The variance for each dataset is a 'within-imputation' variance, while the variance of the MI-estimate also includes a component for 'between-imputation' variance. The variance of the MI-estimate is calculated in two parts. The within-variance ($\bar{U}$) is the estimate of what the variance would have been had the data been complete, and it is calculated as the

mean of the five variances (equation (2) in Section 2.9), in this example, 1.52. The between-variance estimate (B) is the additional variance due to our uncertainty about the missing data. This is calculated by equation (3) in Section 2.9, as 0.57. The between-variance is inflated to adjust for the limited number ($m = 5$) of imputations used and the total variance (T) is the sum of the two components (equation (4) in Section 2.9), giving 2.20. The estimate of the standard error is the square root of this variance (Table 6.5).

Example showing the results from one MCAR dataset

The results from applying the procedure to one set of MCAR data are given in Table 6.5. Table 6.5 gives the (one) MI-estimate of the proportions, combined from the five completed datasets, as above. The MI-estimate of the standard error is an estimate that does not take into account any bias. The 'error' is estimated simply as the difference between the proportion given by the Propensity Score MI-estimate and the complete data proportion. For example, the proportion of women drinking over 14 units of alcohol is estimated as 19.3% from the data completed using SOLAS Propensity Score, 3.7% higher than that in the complete data (15.6%).

Table 6.5: Estimates from SOLAS Propensity Score on a single set of MCAR data. Comparison with estimated proportions over weekly and daily limits from complete data

	Complete data	Propensity Score		
	%	%	SE	Error
Estimated proportions with alcohol consumption over weekly limits				
Women				
>14	15.6	19.3	1.48	3.7
>35	1.1	2.6	0.61	1.5
Men				
>21	33.2	27.5	1.50	−5.7
>50	8.6	7.0	0.89	−1.6
Estimated proportions with alcohol consumption over daily limits				
Women				
>3	38.1	46.2	1.81	8.1
>6	11.5	17.5	1.46	6.0
Men				
>4	63.3	60.9	1.83	−2.4
>8	35.0	32.1	1.54	−2.9

The errors in the estimates of proportions drinking excessively or heavily are consistently large and positive for women, and large and negative for men. The men with missing values apparently drink less and the women more than those whose alcohol consumption was observed,

diluting the difference in drinking between men and women. The results do not reflect the relationship between gender and alcohol consumption.

6.5.4 Discussion

SOLAS Propensity Score has three advantages over the methods used in Section 6.4. Since observed values of alcohol consumption are used to replace the missing values, all imputations are in the valid range. It uses multiple, rather than single, imputation and as such can take account of the uncertainty in the missing data. It avoids the problem posed by the semicontinuous distribution of alcohol consumption.

The fundamental weakness of the method is that it does not preserve relationships between variables. It does not use information from the association among the variables themselves, for example gender and alcohol consumption. It only uses information from covariates that are associated with whether or not the data are missing. A covariate may be related to the variable to be imputed but not to the probability that the variable is missing (as gender in this example). Cases with similar propensity for missingness may be very different in their alcohol consumption. It is possible to define a grouping variable, say gender which would separate models for men and women but this would not solve the problem for other covariates, unless the data is divided into smaller and smaller groups.

The failure to preserve relationships between variables makes the method inappropriate in epidemiological analyses which involve relationships between variables. The imputation algorithm for propensity score originally described by Lavori et al. (1995) was designed for a randomised experiment with repeated measures on the response variable. Some participants dropped out of the study before all the response measurements could be made. The object was to impute the missing responses based on previous response measurements, as well as baseline covariates. Lavori et al. indicate that Propensity Score performs well in this situation.

In the study of alcohol consumption it is not appropriate to avoid the problem of dealing with the semicontinuous distribution of this variable by using Propensity Score. We need to model the associations of variables with alcohol consumption itself.

An alternative solution, proposed by Longford (Longford et al., 2000), is to separate the process of imputing the positive amounts of alcohol consumption (which are approximately log-Normal) and the zeros. This approach is motivated by the ease of working with Normally distributed variables. A similar approach was used by Heitjan and Little (1991) to predict blood alcohol content. Using this approach, we first impute the sign of drinking (0 or 1), according to whether people drank or not, and secondly impute the positive amount of drinking. To implement this approach (without programming the method), we need software which includes procedures both for the imputation of categorical variables and continuous variables. In releases (V 2.0 and subsequently), SOLAS included procedures for multiple imputation of categorical and of continuous variables, and these will be assessed in the next section.

6.6 SOLAS model based procedures

6.6.1 Introduction

SOLAS (v 2.0 and subsequent releases) includes model based procedures for the multiple imputation of categorical variables — 'Discriminant Method' (Section 2.8.4.2.2), and of continuous variables — 'Predictive Model Based Method' (Section 2.8.4.2.1). This enables the analyst to deal with the semi-continuous distribution of alcohol consumption in two separate steps: first imputing the sign of drinking (0 or 1), according to whether people drank or not; and secondly imputing the positive amounts of alcohol consumption (Longford et al., 2000).

6.6.2 Methods

The Predictive Model Based Method is based on linear regression, and assumes that the variable to be imputed is Normally distributed. Whereas the positive amounts of drinking, when log-transformed, are approximately Normally distributed, the presence of the large percentage of zeros presents a problem. For this reason the zero amounts are excluded from the second step by treating them as missing. This is merely a device to avoid the analytic difficulties posed by the semicontinuous distribution. The hypothetical amounts imputed for these zeros are subsequently overwritten by zeros.

Step 1 The sign of drinking — '0' or '1', i.e. whether the respondent drinks or not — is imputed.

Step 2 Any zero amounts in the observed data are set to missing, and the remaining positive amounts are log-transformed. The positive amount of alcohol consumption on each day is then imputed for all missing values, including the zeros that were set to missing.

The imputations for sign and amount are then combined by multiplying the sign by the amount. As a result, values imputed as zero at stage 1 (drank no alcohol) and those observed zeros that were set to missing at step 2 are assigned a zero amount, and values imputed as '1' (drank some alcohol) at step 1 are assigned the amount imputed at step 2. In other words, we impute a hypothetical positive amount for all missing items in the diet diary and for subjects who reported no alcohol consumption. We then overwrite the amount with zero for those who did not drink and also for those whose alcohol consumption was not recorded, but whose sign of drinking was imputed as a zero.

In SOLAS, the Discriminant Method is used to impute signs (step 1) followed by the Predictive Model Based Method, used to impute positive amounts (step 2).

As before, the variables used are given in Section 6.3.2, but the way they are used in each step are specified here.

The independent variables for model of sign (step 1) are as follows:

gender, smoking status, sign of weekly recall, adult social class, CAGE score (coded as binary score 0 or score 1–4), day of week (coded as weekend or weekday), signs of previous days' drinking, days of week of previous days (weekend or weekday).

The independent variables for model of amount (step 2) are:

gender, smoking status, weekly recall amount (log-transformed), CAGE score (coded in three categories 0, 1, and 2–4), adult social class (manual/non-manual), day of the week, amount drunk on other available days (log-transformed), day of week of previous available day.

As for SOLAS Propensity Score, since missing values in covariates must first be imputed, weekly recall and CAGE are first imputed by hot deck conditioning on gender and smoking status (Section 6.5.2).

The SOLAS procedure for multiple imputation proceeds in steps, imputing one variable with missing values at a time, since it works on a monotone structure (see Section 2.7.2 for a discussion of the implications and Section 2.8.4 for the specifics of the procedure). Hence we can specify a different set of covariates for each imputation regression. Some covariates (gender, smoking status, recall, CAGE score, adult social class) are common for any diary day being imputed. Others (the day of the week, the days of the week of the previous days and the previous days' alcohol consumption) depend on the particular day being imputed.

This multiple imputation procedure is set to give five imputations, as before (Section 6.5.2). The five resulting multiple imputation sets are then combined to give the MI-estimate (equation (1) in Section 2.9) and the MI-estimate of the standard error (from equation (4) in Section 2.9). The whole process was repeated eleven times ($n = 11$, Section 6.2.1) for each of the mechanisms of missingness (MCAR, MAR and MNAR). The number of repetitions was limited for practical purposes because the use of the software necessitates manual intervention to import the data sets and to set up each procedure; and the CPU processing time is lengthy: in particular the Predictive Model Based method took around 8 hours on a Pentium II processor and 96MB RAM.

Table 6.6: Estimates from SOLAS Model Based Method for MCAR and MAR data. Comparison with estimated proportions over weekly and daily limits from complete data

	Complete	MCAR					MAR				
	data (%)	BIAS	RMSE	SE	95% CI	*F*	BIAS	RMSE	SE	95% CI	*F*
Estimated proportions with alcohol consumption over weekly limits											
Women											
>14	**15.6**	2.85	2.88	0.47	(1.80, 3.90)	37.0	4.13	4.17	0.62	(2.75, 5.51)	44.5
>35	**1.1**	0.77	0.78	0.15	(0.44, 1.10)	26.3	1.39	1.42	0.28	(0.77, 2.01)	23.9
Men											
>21	**33.2**	4.23	4.28	0.65	(2.78, 5.68)	42.2	5.83	5.93	1.07	(3.45, 8.21)	29.4
>50	**8.6**	1.34	1.40	0.38	(0.49, 2.19)	12.8	2.46	2.53	0.61	(1.10, 3.82)	16.2
Estimated proportions with alcohol consumption over daily limits											
Women											
>3	**38.1**	3.02	3.07	0.57	(1.75, 4.29)	28.3	3.77	3.82	0.64	(2.34, 5.20)	34.1
>6	**11.5**	3.89	3.91	0.37	(3.07, 4.71)	108.8	5.73	5.79	0.82	(3.90, 7.56)	49.5
Men											
>4	**63.3**	2.34	2.45	0.72	(0.74, 3.94)	10.5	3.67	3.69	0.44	(2.69, 4.65)	70.8
>8	**35.0**	3.77	3.88	0.88	(1.81, 5.73)	18.3	5.48	5.56	0.89	(3.50, 7.46)	38.3

Percentage points of $F_{1,10}$: $P = 0.05$ $F = 4.96$, $P = 0.01$ $F = 10.0$, $P = 0.001$ $F = 21.0$

6.6.3 Results

Table 6.6 gives the estimated biases and their standard errors from the 11 repetitions of the SOLAS procedures for the MCAR and MAR data. Although some of the biases for MCAR are relatively small compared with the other methods (for example, MVR in Table 6.1 and SPSS Regression and EM estimates over daily limits in Table 6.3), they are consistently positive for all the estimates for both MAR and MCAR. The standard errors are small relative to the bias and the lower limits of the 95% confidence intervals are all above zero indicating that the biases did not arise by chance. Using the F-test each bias is statistically significant, generally at the level $P < 0.001$ (and $P < 0.01$ for the proportion of men drinking more than 50 units over the week), despite the small number of repetitions. Also the positive bias for the MAR data is consistently greater than that for the corresponding estimate for the MCAR data. These results indicate that there is either a problem with the algorithms used by the methods or in the programming of those algorithms. The next section investigates the source of this problem.

6.6.4 Further diagnostic tests of the SOLAS Discriminant Method

6.6.4.1 The origin of the positive bias produced by SOLAS under MCAR

The positive bias in the imputations produced by SOLAS under MCAR was found in the imputation of the sign of drinking: that is, the method resulted in an overestimation in the proportion of people who drank on any one day. Figure 6.3 shows the proportion of people drinking (positive sign) on each day of the week in those whose drinking was fully observed and those who had some (or all) of their diary day records imputed. Although the pattern by day of the week is preserved, the imputation of positive sign is positively biased for men and women

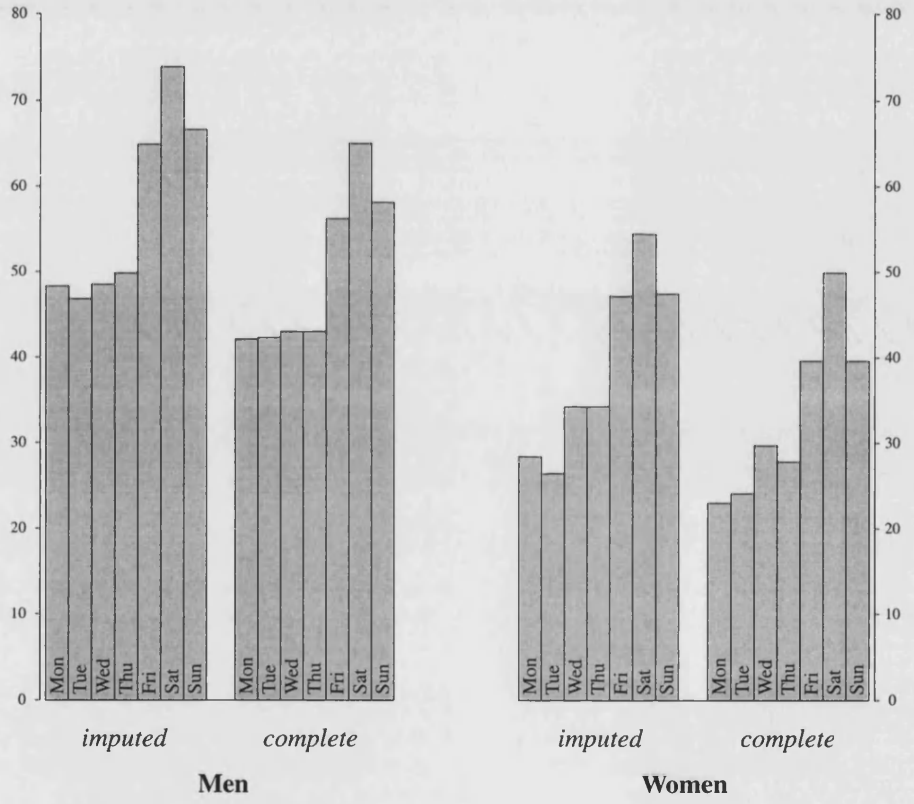


Figure 6.3. Proportion of men and women drinking on each day of the week amongst those with complete data compared with those who had some (or all) of their diary day records imputed using the SOLAS Discriminant Method

6.6.4.2 Theoretical problems: assumptions of SOLAS Discriminant Method

The SOLAS procedure Discriminant Method is designed to impute categorical variables with either categorical or continuous covariates. However, the method assumes that the covariates (X) have a multivariate Normal distribution for each value of the variable to be imputed (Y) (see Section 2.8.4.2.2 (v)). The assumption of multivariate Normality can, theoretically, only be applied to continuous covariates (which are multivariate Normal) and not to categorical covariates. However, as for all statistical methods, in practice the application may be robust to

departures from its theoretical assumptions. The robustness of the method in practice will depend on the estimator, the relationship between the variable to be imputed and the categorical covariates.

Effect of the multivariate Normal assumption in practice

The imputed sign of drinking was found to be positively biased whenever the analysis included as covariate the binary variable which indicated whether the person had drunk at all during the previous week (sign of the previous week's drinking). As an example, the relationship between this indicator and the third day's drinking is shown in the Table 6.7.

Table 6.7: Relationship between sign of weekly recall and sign of alcohol consumption on the third day of the diary for diary completers

Any alcohol consumption on the third diary day	Any alcohol consumption in the previous week		
	No (sign=0)	Yes (sign=1)	Total
No (sign=0)	423	811	1234
	94.4%	56.3%	62.9%
Yes (sign=1)	25	703	728
	5.6%	46.4%	37.1%
Total	448	1514	1962
Row %	22.3%	75.6%	

Most people drink, and those who reported not drinking in the last week are unlikely to report drinking in the diary week. A proportion of who report not drinking in the last week will be non-drinkers and those who do drink are more likely to be occasional drinkers. In other words no alcohol consumption in the previous week (sign = 0) is highly predictive of no (sign = 0) alcohol consumption on any one day. This relationship leads to the bias resulting from the SOLAS method becoming apparent: this is demonstrated with simulated data below.

6.6.4.3 Using simulated data to test SOLAS Discriminant Method

Simulated dataset 1

To investigate how the bias arises, a simpler simulated dataset was used with one binary covariate (X) and one binary variable to be imputed (Y), related in a similar way to the sign of drinking in the weekly recalled drinking and drinking on a diary day, as shown in Table 6.8.

After MCAR deletion of approximately 38% of Y, we get the results in Table 6.9:. Y was then imputed using covariate X only using the SOLAS procedure 'Discriminant Method'. The overall results including observed and imputed values, averaged over the 5 imputed data sets, are given in Table 6.10. Using imputed data sets from the application of the SOLAS Discriminant Method, the estimation of the proportion of cases with $Y = 1$ is 57.2% (Table

6.10). This is clearly positively biased compared with the known proportion of 48.9% for the complete 2002 cases (Table 6.8).

To see how this bias arises, we look at the imputed values in isolation. Table 6.11 gives the values of Y that were set to missing by the MCAR deletion, obtained by subtracting the number in Table 6.9 from those in Table 6.8. These are the values of Y to be imputed. However, the SOLAS software produced the sets of imputed values for Y given in Table 6.12.

The number of imputations of $Y = 1$ when $X = 0$, ($n = 0$, Table 6.12) is too low, although this has little effect because we had only one instance of this in the missing cases; whereas the number of imputations of $Y = 1$ when $X = 1$ are too high (Table 6.12: 542, 516, ... , mean=533.4). This has quite a dramatic effect because there were 367 instances in the missing cases (Table 6.11). The log output for SOLAS gives the probabilities of imputing each sign of Y for each case. These vary for each data set since they have a random perturbation added, but for example for one data set $P(Y = 1 \mid X = 1) = 0.81$; $P(Y = 1 \mid X = 0) = 2.0 \times 10^{-18}$. These probabilities, generated by the software, are clearly biased compared with the observed proportions of 4.1% and 55.7% (Table 6.9) from which they are estimated. The probabilities for each data set are positively biased for $Y = 1$ when $X = 1$ and negatively biased for $Y = 1$ when $X = 0$. The positively biased results for $Y = 1$ follow since there is a higher proportion of cases with $X = 1$ than $X = 0$.

The final step was to check whether the bias produced by using the SOLAS Discriminant Method could be the result of a programming error. The results given by the software and presented above were checked by calculating the conditional probabilities for imputation of $Y \mid X = 0$ and $Y \mid X = 1$ manually using the SOLAS algorithm as specified in Section 2.8.4.2.2. Details of these manual calculations are given in Appendix 3, SOLAS Discriminant Method. The results of the manual calculation show that $P(Y = 1 \mid X = 0) = 3.3 \times 10^{-18}$ (equation (13) in Appendix 3) and $P(Y = 1 \mid X = 1) = 0.82$ (equation (15) in Appendix 3), agreeing with the output from the software. Using these calculated probabilities the predicted numbers of imputations for $Y = 1$ when $X = 0$ would be 0, and for $Y = 1$ when $X = 1$ would be 533.5, agreeing with the SOLAS output averaged over the 5 data sets (Table 6.12). From this manual calculation we can conclude that the algorithm used by SOLAS is correctly implemented and it is the method itself, which leads to biased results.

The biases result from the application of the Normal distribution assumption when the conditional distribution of $X \mid Y$ is skewed, as $X \mid Y = 1$ is in this example. The Normal distribution uses the estimated variance of $X \mid Y$ which is very small when the expected value of X is near 1 (as in this case, see equation (7) in Appendix 3), or 0. When such a very small value is in the divisor (as in equation (11) in Appendix 3) it leads to positive bias; when it is in the exponent of the Normal density function (equation (10) in Appendix 3) it leads to negative bias.

This is an extreme example as the distribution of $X \mid Y$ is very skewed, i.e. the probability that $X = 1$ for any value of Y is far from 0.5. In general the use of the Normal distribution will

Table 6.8: Complete simulated dataset 1

Y	*X*= 0	*X*= 1	Total
Y= 0	262	762	1024
	97.0%	44.0%	51.1%
Y= 1	8	970	978
	3.0%	56.0%	48.9%
Total	270	1732	2002
(row %)	13.5%	86.5%	

Table 6.9: Incomplete simulated dataset 1, after MCAR deletion of *Y*

Y	*X*= 0	*X*= 1	Total
Y= 0	164	479	643
	95.9%	44.3%	51.3%
Y= 1	7	603	610
	4.1%	55.7%	48.7%
Total	171	1082	1253
(row %)	13.6%	86.4%	

Table 6.10: Dataset 1 completed by MI for missing values of *Y* using the Discriminant method

Y	*X*= 0	*X*= 1	Total
Y= 0	263	593.6	856.6
	97.4%	34.3%	42.8%
Y= 1	7	1138.4	1145.4
	2.6%	65.7%	57.2%
Total	270	1732	2002

Table 6.11: Dataset 1, values of *Y* to be imputed

Y	*X*= 0	*X*= 1	Total
Y= 0	98	283	381
Y= 1	1	367	368
Total	99	650	749

Table 6.12: Multiple Imputation of Y values in dataset 1 using SOLAS 'Discriminant method'

	X =		*X* =		*X* =		*X* =		*X* =		*X* =	
Y	0	1	0	1	0	1	0	1	0	1	0	1
Y= 0	99	108	99	134	99	120	99	115	99	106	99	116.6
Y= 1	0	542	0	516	0	530	0	535	0	544	0	533.4
Total	99	650	99	650	99	650	99	650	99	650	99	650
Imputation	1		2		3		4		5		Mean	

increase the bias in the estimated probabilities (of $Y \mid X$) the greater the skew, i.e. the further from 50% are the proportions of $X = 0$ or $X = 1$ for a given value of Y. However the nearer these proportions are to 50% the poorer X is as a predictor of Y: so the less informative X is as a covariate for Y!

In some cases, the bias resulting from the Normal distribution assumption may not be evident in the marginal distribution of Y, as illustrated by the second simulated dataset.

Simulated dataset 2

In the second simulated dataset the relationship between X and Y was 'balanced', so that the proportion of $Y = 1 \mid X = 0$ is similar to the proportion of $Y = 0 \mid X = 1$. This is shown in the cross-classification of X and Y given in Table 6.13. After MCAR deletion of approximately 38% of Y, we get the results in Table 6.14.

Table 6.13: Complete simulated dataset 2

Y	*X*= 0	*X*= 1	Total
Y= 0	762	262	1024
	74.8%	26.7%	51.1%
Y= 1	257	721	978
	25.2%	73.3%	48.9%
Total	1019	983	2002
(row %)	50.9%	49.1%	

Table 6.14: Incomplete simulated dataset 2, after MCAR deletion of Y

Y	*X*= 0	*X*= 1	Total
Y= 0	479	160	639
	74.5%	26.2%	51.0%
Y= 1	164	450	614
	25.5%	73.8%	49.0%
Total	643	610	1253
(row %)	51.3%	48.7%	

When the SOLAS Discriminant Method is applied to this data the MI-estimate for $Y = 1$ is 48.9%, and is clearly unbiased relative to that in the complete data (48.9%, Table 6.13). For given values of X, however, the MI-estimate for the proportion of $Y = 1 \mid X = 0$ is 23.3% and for $Y = 1 \mid X = 1$ it is 75.6%. Compared with the corresponding observed proportions in the incomplete dataset in Table 6.14 (25.3% and 73.8%), these MI-estimates are negatively and positively biased (respectively), as would be predicted from the above argument (for Simulated dataset 1). This is because the negative bias in the proportion of $Y = 1 \mid X = 0$ (23.3%) and the positive bias in $Y = 1 \mid X = 1$ (75.6%) cancel each other out. However, even though this may

give unbiased results for the marginal distribution of Y, the relationship between X and Y is not preserved in the imputed data sets because of the biases in the estimated probabilities for given values of X. For example, the known odds ratio, for $Y = 1$, comparing $X = 1$ with $X = 0$, is 8.16 (calculated from the complete dataset, Table 6.14 by $(721 \times 762)/(262 \times 257)$). The same odds estimated from the imputed data sets is 10.06. Hence although marginal estimates may be robust to the bias resulting from the SOLAS Discriminant Method, estimates of the relationships between the covariate and the variable to be imputed may not be.

6.6.4.4 Alternative procedure for imputing a categorical variable using only categorical covariates

Another procedure for imputing categorical variables is Schafer's programme 'CAT' (Section 2.8.5). CAT uses a loglinear model, which is more appropriate for categorical data than the model used by the SOLAS Discriminant Method. CAT was used to generate imputations using the incomplete simulated datset1 (Table 6.9). Using the multiple datasets produced by CAT, the MI-estimate of the proportion of cases with $Y = 1$ is 48.8%, which is unbiased compared with the known proportion of 48.9% for the complete dataset (Table 6.8). The MI-estimate of the proportion of cases with $Y = 1 \mid X = 0$ is 4.4% (from CAT) compared with 4.1% for the observed values in the incomplete dataset (Table 6.9); the corresponding estimate for $Y = 1 \mid X = 1$ is 55.8% (from CAT) compared with the observed 55.7% (Table 6.9). The results are unbiased and the relationship between X and Y is preserved.

6.6.5 Summary

In Section 6.5 it was found that imputation of the variable (Y) with missing values by using the Propensity Score procedure, which models the missingness (R) of Y, was not appropriate, since it does not preserve relationships between the variable Y and covariates X. In order to preserve such relationships it is necessary to model the variable with missing data (Y) directly. The procedures available in the software SPSS to deal with missing data by modelling the variable with missing data were reviewed in Section 6.4. Besides the specific disadvantages of these SPSS procedures, they (and other procedures for modelling a continuous variable Y in relation to covariates X) assume that Y is Normally distributed. In this application the variable Y is alcohol consumption, which has a semi-continuous distribution, and it has been shown that the SPSS procedures are not robust to this semi-continuous distribution.

It is possible to deal with the semi-continuous distribution of alcohol consumption in two separate steps: first imputing the sign of drinking (0 or 1), according to whether people drink or not; and secondly imputing the (continuous) positive amounts of alcohol consumption. The first step requires a procedure for imputing categorical variables; the second a procedure for imputing continuous variables. Since the software package SOLAS includes both types of procedure (the 'Discriminant Method' for imputing categorical variables and the 'Predictive Model Based Method' for continuous variables), it could be used to implement the two-step method.

However, when the SOLAS procedures were used the resulting estimates were biased when the data was MCAR. In this section (6.6), it has been shown that the problem arises in the SOLAS 'Discriminant Method'. This procedure is not appropriate when the covariates (X) are categorical, because it assumes that the covariates have a multivariate Normal distribution for each value of the variable to be imputed (Y). It has been shown that the procedure is not robust to these assumptions when the covariates are binary and the conditional distribution of $X|Y$ is skewed (simulated dataset 1). Furthermore, it is not suitable for epidemiological applications because, even when the estimated marginal distributions of Y are unbiased, the relationship between X and Y may not be (as shown by simulated dataset 2). Where all covariates are categorical, as is the case for the imputation of the sign of drinking in the diary, it is more appropriate to use a loglinear model. A procedure which implements the loglinear model is Schafer's CAT. This procedure was shown to produce unbiased MI estimates for the skewed data (simulated dataset 1). In Section 6.7 Schafer's software will be used on the simulated alcohol data.

6.7 Schafer's procedures

6.7.1 Introduction

This section explores Schafer's software for multiple imputation. In the previous section a two-step approach separating the imputation of the sign of drinking and the positive amount of drinking (defined in Section 6.6.2), was used to avoid the problem posed by the semi-continuous distribution of alcohol consumption (Section 2.4). This approach requires procedures for imputing categorical and continuous variables. Schafer's software (described in detail in Section 2.8.5) includes a package for use with categorical data (CAT), and one for use with continuous data (NORM), so the two-step approach can be implemented using this software. Schafer's CAT has been shown to give appropriate results with skewed binary data as a covariate (Section 6.6.4.4) and NORM with correlated Normally distributed data (Section 6.4.4). Schafer has also produced MIX, a package specifically designed for a mixture of categorical and continuous variables.

The fundamental difference between the approaches used in SOLAS and Schafer's software is that the former proceeds in a step-by-step fashion, imputing each variable one at a time using separate regressions, whereas the latter imputes all missing data simultaneously in a multivariate regression (Section 2.7.2). In this respect, the latter approach has the theoretical advantage of greater efficiency because it can maintain the relationships among all the variables in the model simultaneously. However, the step-by-step approach used by SOLAS has a practical advantage in the current application. The reduction to a series of single-variable imputations allows 'tremendous modeling flexibility' (Rubin, 2000): that is, a different set of covariates can be used for each variable to be imputed. This means that the day of the week (Monday, Tuesday, ...) can be associated individually with each diary day. The ability to identify and take account of the day of the week is important in the context of estimating alcohol consumption from a week's diary because of the variation of drinking over the days of the week (or the pattern of alcohol

consumption) (Section 6.2.2). The method using Schafer's software is refined to improve the representation of this pattern.

6.7.2 Methods

Accounting for the day of the week

Representing the pattern of drinking over the diary week depends on being able to identify the day of the week of each diary day. However, the diary started on different days of the week (Section 3.2.3). The data, as it is ordered by diary day, has a monotone pattern of missingness (Section 2.7.1) and this monotone pattern is preserved in the simulated data (Section 6.2.3). With the data in a monotone structure, SOLAS Predictive Model Based Method proceeds in a step by step fashion, imputing each variable (diary day's alcohol consumption) one at a time using separate regressions, starting with the first diary day, which has the lowest proportion of missingness, and proceeding day by day through the diary week. This means that using SOLAS different covariates may be entered into the regression equation at each step (for each day of the diary), and we are able to include the day of the week separately for each diary day. In contrast, Schafer's software imputes all missing data simultaneously in a multivariate regression. Using Schafer's software we can account for the day of the week only by including the day of the week factor once: in this case we specify the day of the week on which the diary starts (*startday*) as a covariate (Method 1). As a consequence, the Method can condition on the day of the week of the first diary day only. It cannot associate the days of the week for respondents who start their diaries on different days. For example, for all those who start their diary on a Monday, their subsequent days can be associated by their sequence (their second days are Tuesday, and so on), but for someone who starts their diary on a Wednesday, Monday, being the sixth diary day, cannot be associated with the Monday for the first group.

An alternative way of accounting for the pattern of alcohol consumption is considered by reordering the diary records so that they are all in weekday order (Monday through to Sunday). With the data in this structure, Schafer's software can take into account the pattern of alcohol consumed on each day of the week using all the available data, since each variable then represents a day of the week (Method 2). With startday as a factor, Method 2 can also take into account the effect of starting the diary on different days of the week. The importance of the diary day order is in the difference in the data collection in the first two days (retrospectively completed by the nurse at the interview) versus the following 5 days (completed by the respondent themselves during the following days). However the effect of these ways of collecting the data has been shown to be negligible (Section 5.6). The inclusion of 'startday' as a factor increases the number of parameters to be estimated seven-fold, and so increases the segmentation of the data, so that less data is available for the estimation of each parameter. Method 3 excludes the factor 'startday'.

The procedure MIX presents the additional possibility of imputing the sign and the positive amounts simultaneously. Using MIX we can impute the sign and amount of alcohol

consumption in one step (Method 4). With Method 4 the values of the signs of alcohol consumption can influence the estimation and the imputation of the amounts of alcohol consumed (although it cannot associate the sign and the amount for the same day of the week). This was thought to be important because the pattern of signs of drinking are associated with the amount drunk: for example, the more days people drank the more they tended to drink on each day (Table 5.5). Hence a method using MIX may better reflect the pattern of drinking than methods that used a two-step approach.

Assessing the pattern of alcohol consumption

The pattern of alcohol consumption is assessed using bar graphs of the signs and amounts of alcohol consumption over the days of week, as detailed in Section 5.5. Here, the pattern in the imputed values using MCAR data is compared with the *complete data.* The preservation of pattern can be assessed by comparing the outline shape of the bar graphs. Since the data are MCAR the *observed* values available to the imputation method (the records not deleted by applying the MCAR mechanism) will be representative of the *complete data* values. For a good method, the (known) pattern in the completed data should be reflected in the pattern of the imputed values. As an example the graphs presented are of the signs and amounts of alcohol consumption on each day of the week for men derived from one set of imputations (the first) produced using one of the MCAR datasets (the final one in the repetition of the process).

Variables used in the imputation model

Apart from the issue of the day of the week, the models use essentially the same covariates as with SOLAS (Section 6.6.2). The same two-step procedure is used, except when using MIX. Using Schafer's software all the variables are entered, and missing data on any are imputed simultaneously, so there is no need to impute the 'covariates' first. The variables used for the imputation of sign of alcohol consumption in CAT are: sign of alcohol consumption on the available seven diary days, gender, smoking status, sign of weekly recall, adult social class, CAGE (coded in the two categories 0–1 and 2–4), *startday* (day of the week of the first diary day). The variables used for the imputation of positive quantities in NORM are: the log-transformed positive alcohol consumption in grams on the available seven diary days, gender, smoking status, weekly recall amount (in units, log-transformed after adding unity), CAGE score, adult social class, and startday . NORM treats categorical covariates as continuous. In Method 3, 'startday' is omitted. In Method 4 the signs and quantities are imputed in one step using MIX. The variables entered as categorical are: sign of alcohol consumption on the available seven diary days, gender, smoking status, sign of weekly recall, adult social class, CAGE (0–1/2–4). The variables entered as continuous are: log-transformed positive alcohol consumption in grams on the available seven diary days, and weekly recall amount (in units, log-transformed after adding unity).

Imputation and Data Augmentation

Schafer's procedures comprise four steps: a preliminary computation step, an EM step, a Data Augmentation (DA) step and an imputation step (see Section 2.8.5 for details). For Methods 1-4 in the EM step a saturated model was used for CAT and MIX. The saturated model includes all of the interactions among the categorical variables. This approach is used because the interactions between the categorical variables were found to be important for predicting alcohol consumption (see Sections 5.3 & 5.5). The imputation step can be used to give multiple imputations from the sampling distribution derived from the EM directly, without using the DA step. The DA step uses MCMC (Markov Chain Monte Carlo, Section 2.8.5) sampling of the parameter values, and is used to represent additional uncertainty about the estimated parameters. For simplicity, the DA step was omitted in the development of the method. Methods 1–4 do not include DA. Five independently drawn imputations were obtained, each by repetition of the imputation step. The DA step would not run using MIX (Method 4), because too many parameters had to be estimated (a multivariate Normal distribution is associated with each combination of categories, Section 2.8.5). It was necessary to simplify the model for the DA step to an independence model with just two categorical variables in addition to the seven signs of alcohol consumption. Gender and the sign of weekly recall were chosen since they were considered to be the most important of the categorical variables (Method 5).

As in Sections 6.5 and 6.6, the whole process results in a set of five imputations, and the five resulting multiple imputation sets are combined as described in Section 2.9. Methods 1–5 are each applied to 30 repetitions of the simulation process (n = 30, Section 6.2.1) using the MCAR mechanism, and Methods 2–4 are applied using the MAR mechanism of missingness.

In summary, the approaches using SchaferÂs software are as follows:

METHOD 1. Using CAT and NORM: data ordered by diary day, with the day of the week of the first diary day ('startday') as a factor

METHOD 2. Using CAT and NORM: data ordered by day of the week, with 'startday' as a factor

METHOD 3. Using CAT and NORM: data ordered by day of the week, without 'startday' as a factor

METHOD 4. Using MIX only: data ordered by day of the week, without 'startday' as a factor

METHOD 5. MIX with DA, using an independence model with the categorical variables restricted to gender, sign of recall and signs of the seven days' drinking.

6.7.3 Results

The results for all the Methods, identified with their Method number, are given in Table 6.15 (for the MCAR datasets), and Table 6.16 (for the MAR datasets).

The biases are all small (in absolute value) compared with those using SOLAS model based procedures (Table 6.6). When the data is MCAR (Table 6.15) the biases are relatively large compared with the standard errors only for Method 5. The biases for most of the estimates using Method 5 have 95% confidence intervals excluding zero (shown in bold) and are statistically significant at the 5% level. The evidence suggests that the estimated proportions of men drinking over weekly limits are negatively biased, and that those of women drinking over daily limits are positively biased. Since it was biased when the data was MCAR, Method 5 was not applied to the MAR data. The hypothesis of no bias is rejected only for Methods 1–3 under MCAR (Table 6.15) and Methods 2–4 under MAR (Table 6.16) for one estimate: the proportion of women drinking heavily (more than 6 units on any one day). When Method 3 is applied to MAR data (Table 6.16) the biases are generally lower than those obtained using Method 2. For Method 4, the biases are all lower than for any other Method and none of them are statistically significant (at the 5% level) under MCAR.

[Text continues after Tables 6.15 and 6.16 below]

Table 6.15: Methods using Schafer's procedures to estimate proportions over weekly and daily limits for MCAR data: comparison with estimates from complete data

Comparison of estimated drinking over weekly limits for MCAR data

			METHOD 1 CAT and NORM Diary day order + startday					METHOD 2 CAT and NORM Diary day order + startday					METHOD 3 CAT and NORM Weekday order no startday					METHOD 4 MIX Weekday order no startday					METHOD 5 MIX with DA Restricted independence model				
Complete data																											
f	%	BIAS	RMSE	SE	*F*	95% CI	BIAS	RMSE	SE	*F*	95% CI	BIAS	RMSE	SE	*F*	95% CI	BIAS	RMSE	SE	*F*	95% CI	BIAS	RMSE	SE	*F*	95% CI
>14	15.6	−0.21	0.52	0.48	0.18	(−1.19,0.77)	−0.20	0.60	0.57	0.13	(−1.37,0.97)	−0.11	0.56	0.55	0.04	(−1.23,1.01)	0.11	0.71	0.70	0.02	(−1.32,1.54)	−1.29	1.87	0.44	8.72	**(−2.19,−0.39)**
>35	1.1	0.26	0.32	0.19	1.81	(−0.13,0.65)	0.23	0.30	0.19	1.47	(−0.16,0.62)	0.27	0.33	0.18	2.43	(−0.10,0.64)	0.14	0.20	0.14	0.99	(−0.15,0.43)	−0.12	0.04	0.17	0.56	(−0.47, 0.23)
m																										
>21	33.2	−0.48	0.91	0.77	0.39	(−2.05,1.09)	−0.86	1.09	0.68	1.61	(−2.25,0.53)	−0.61	0.83	0.56	1.17	(−1.76,0.54)	−0.38	0.81	0.71	0.30	(−1.83,1.07	−3.85	15.39	0.74	27.30	**(−5.36,−2.34)**
>50	8.6	−0.51	0.61	0.34	2.26	(−1.21,0.19)	−0.47	0.59	0.35	1.75	(−1.19,0.25)	−0.13	0.37	0.35	0.15	(−1.85,0.59)	−0.42	0.63	0.47	0.79	(−1.38,0.54)	−1.90	3.76	0.40	22.26	**(−2.72,−1.08)**

Comparison of estimated drinking over daily limits for MCAR data

			METHOD 1 CAT and NORM Diary day order + startday					METHOD 2 CAT and NORM Diary day order + startday					METHOD 3 CAT and NORM Weekday order no startday					METHOD 4 MIX Weekday order no startday					METHOD 5 MIX with DA Restricted independence model				
Complete data																											
f	%	BIAS	RMSE	SE	*F*	95% CI	BIAS	RMSE	SE	*F*	95% CI	BIAS	RMSE	SE	*F*	95% CI	BIAS	RMSE	SE	*F*	95% CI	BIAS	RMSE	SE	*F*	95% CI
>3	38.1	−0.33	0.81	0.74	0.21	(−1.84,1.18)	−0.15	0.84	0.83	0.03	(−1.85,1.55)	−0.58	0.79	0.54	1.13	(−1.68,0.52)	0.05	0.85	0.84	0.00	(−1.67,1.77)	2.56	7.49	0.97	6.99	**(0.58, 4.54)**
>6	11.5	1.19	1.25	0.39	9.25	**(0.39,1.99)**	1.25	1.35	0.52	5.84	**(0.19,2.31)**	1.21	1.31	0.51	5.68	**(0.17,2.25)**	0.93	1.07	0.52	3.19	(−0.13,1.99)	1.73	3.32	0.57	9.12	**(0.56, 2.90)**
m																										
>4	63.3	−0.89	1.29	0.93	0.92	(−2.79,1.01)	−1.22	1.42	0.72	2.93	(−2.69,0.25)	−1.07	1.26	0.66	2.58	(−2.42,0.28)	−0.55	0.89	0.70	0.61	(−1.98,0.88)	−0.65	1.36	0.97	0.45	(−2.63, 1.33)
>8	35.0	−0.45	1.01	0.90	0.25	(−2.29,1.39)	−0.54	0.92	0.75	0.52	(−2.07,0.99)	−0.66	0.90	0.61	1.16	(−1.91,0.59)	0.00	0.83	0.83	0.00	(−1.70,1.70)	−1.33	2.41	0.80	2.77	(−2.97, 0.31)

Percentage points of $F_{1,29}$: $P = 0.05$ $F = 4.18$, $P = 0.01$ $F = 7.60$, $P = 0.001$ $F = 13.4$

Table 6.16: Methods using Schafer's procedures to estimate proportions over weekly and daily limits from MAR data: comparison with proportions estimated from complete data

Estimated proportions with alcohol consumption over weekly limits: MAR data

	Complete data	METHOD 2 CAT and NORM Weekday order with startday					METHOD 3 CAT and NORM Weekday order no startday					METHOD 4 MIX Weekday order no startday				
Women	**%**	BIAS	RMSE	SE	95% CI	*F*	BIAS	RMSE	SE	95% CI	*F*	BIAS	RMSE	SE	95% CI	*F*
>14	**15.6**	−1.42	1.53	0.57	(−2.59, −0.25)	6.30	−0.58	0.81	0.57	(−1.75,0.59)	1.05	−0.03	0.67	0.67	(−1.40, 1.34)	0.00
>35	**1.1**	0.16	0.24	0.77	(−1.41 , 1.73)	0.91	0.40	0.46	0.23	(−0.07,0.87)	2.97	0.09	0.25	0.23	(−0.38, 0.56)	0.16
Men																
>21	**33.2**	−2.17	2.29	0.75	(−3.70, −0.64)	8.47	−0.97	1.22	0.75	(−2.50,0.56)	1.68	−0.36	0.78	0.69	(−1.77, 1.05)	0.27
>50	**8.6**	−1.32	1.35	0.31	(−1.95, −0.69)	18.43	−0.54	0.70	0.45	(−1.46,0.38)	1.40	−1.08	1.20	0.52	(−2.14,−0.02)	4.38

Estimated proportions with alcohol consumption over daily limits: MAR data

	Complete data	METHOD 2 CAT and NORM Weekday order with startday					METHOD 3 CAT and NORM Weekday order no startday					METHOD 4 MIX Weekday order no startday				
Women	**%**	BIAS	RMSE	SE	95% CI	*F*	BIAS	RMSE	SE	95% CI	*F*	BIAS	RMSE	SE	95% CI	*F*
>3	**38.1**	−0.60	1.04	0.84	(−2.32, 1.12)	0.51	−0.43	0.75	0.61	(−1.68,0.82)	0.48	0.93	1.19	0.75	(−0.60, 2.46)	1.54
>6	**11.5**	1.47	1.55	0.51	**(0.43, 2.51)**	8.18	1.86	1.92	0.47	**(0.90,2.82)**	15.54	1.83	1.91	0.54	**(0.73, 2.93)**	11.44
Men																
>4	**63.3**	−1.47	1.56	0.52	(−2.53, −0.41)	8.02	−0.89	1.21	0.81	(−2.55,0.77)	1.23	−0.16	0.65	0.63	(−1.45, 1.13)	0.07
>8	**35.0**	−0.94	1.07	0.52	(−2.00, 0.12)	3.29	−0.61	0.88	0.63	(−1.90,0.68)	0.93	0.27	0.75	0.70	(−1.16, 1.70)	0.15

Percentage points of $F_{1,29}$: $P = 0.05$ $F = 4.18$, $P = 0.01$ $F = 7.60$, $P = 0.001$ $F = 13.4$

Figure 6.4 shows the pattern of signs of consumption for men for imputation Methods 1–4 and Figure 6.5 the pattern of log-transformed positive amounts. The patterns for Model 5 using Data Augmentation with a restricted independence model are given in Figure 6.6. For each Method the pattern in the complete data is given on the right hand side for comparison with the imputed pattern on the left. The main feature of the pattern of signs for the complete data is that the weekdays are clearly differentiated from weekend days (Figure 6.4). As reported in Section 5.5, respondents were more likely to drink at the weekends (Friday, Saturday and Sunday) than during the week. The pattern in the amounts for complete data (Figure 6.5, right hand side) is similar in this respect, except that on average men drink less on Sunday than on Friday and Saturday. The difference between the weekdays and weekends is blurred in the imputations from Method 1, particularly for the positive amounts imputed: the pattern of alcohol consumption in the complete data is not preserved by Method 1. For Method 2, the pattern of drinking in the imputations is an improvement over Method 1: the popularity (Figure 6.4) and level of drinking (Figure 6.5) on Saturdays is better reflected in these imputations, although Friday and Sunday drinking are still underestimated. The pattern in the imputed signs is better preserved by Method 3 (Figure 6.4) than Method 1 or Method 2: the weekend being clearly differentiated from the weekdays. Whilst the amounts for Method 3 (Figure 6.5) are lower than the known values, particularly for consumption on Sunday, those for Friday and Saturday are differentiated from the weekdays. The pattern in the imputations for Method 4 preserves the pattern of signs better than the other Methods, but the amounts are rather low. The patterns for Method 5 are given in Figure 6.6. The imputations from Method 5 are poorer at reflecting the pattern of drinking than any of the Methods 2, 3 or 4.

[Text continues after Figures 6.4–6.6 below]

Figure 6.4: Proportion of men drinking on each day of the week for Methods 1–4
The graph for each Method consists of two parts: the *left hand* side gives the imputed values from one set of imputations, and the *right hand* side gives the complete data

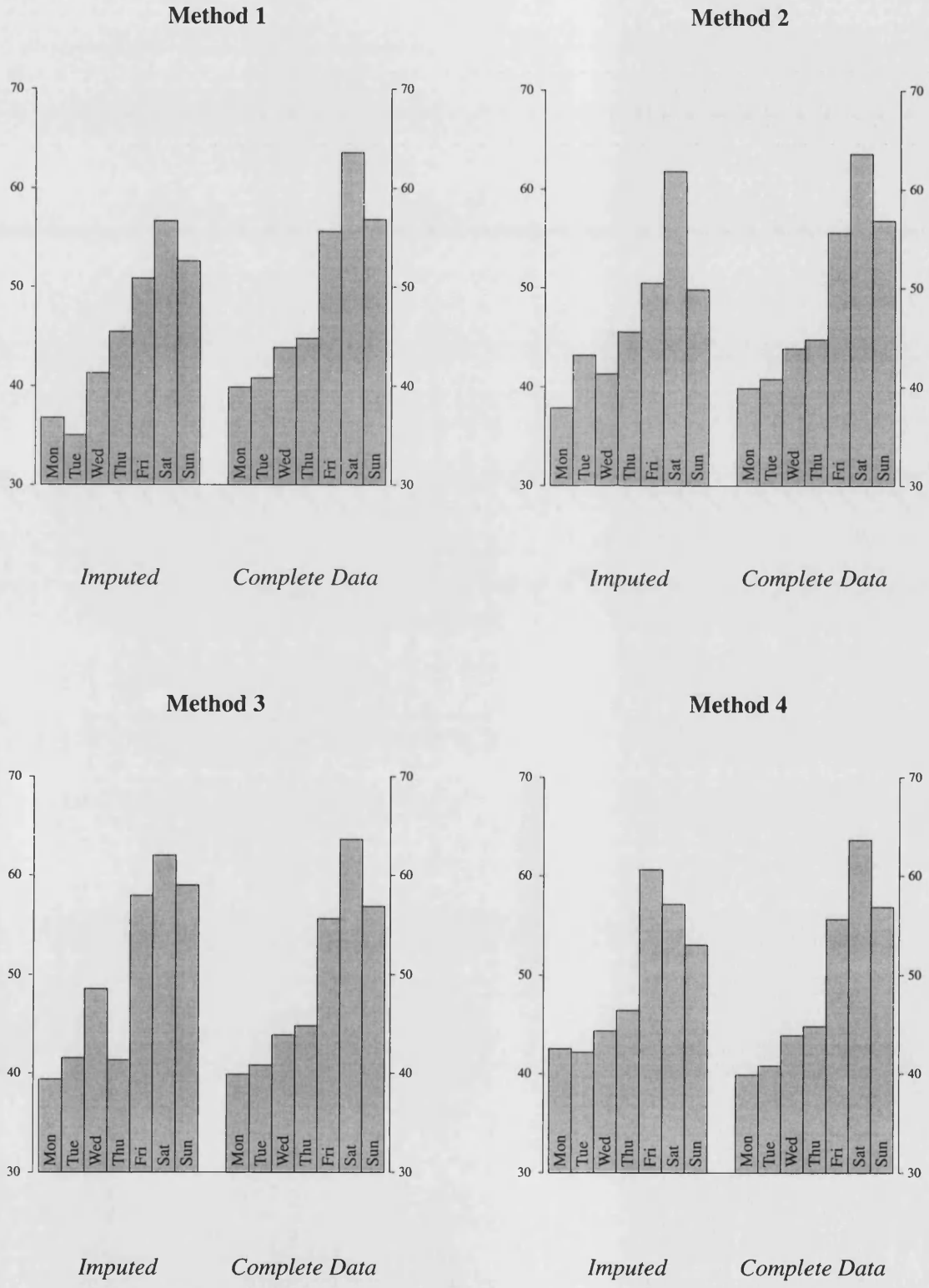
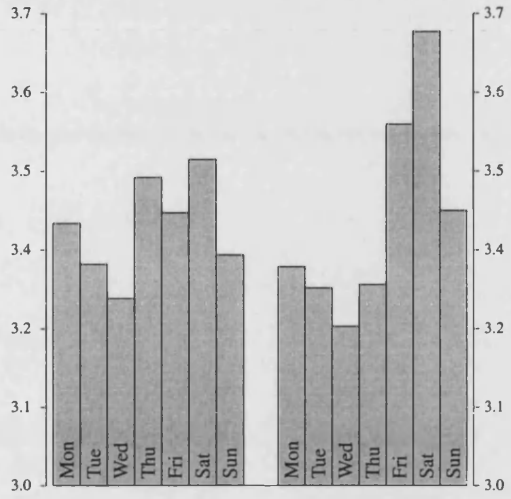


Figure 6.5: Mean of log-transformed positive amounts drunk by men on each day of the week for Methods 1–4

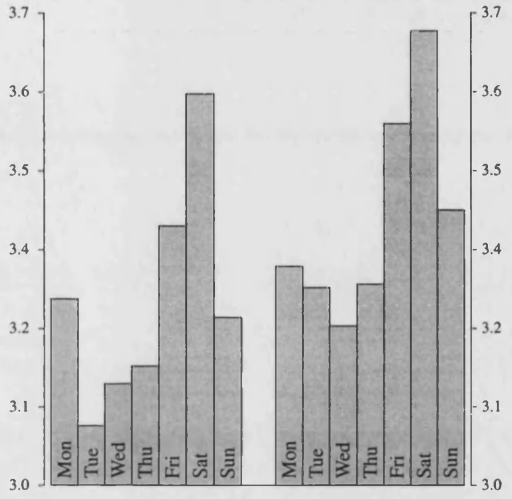
The graph for each Method consists of two parts: the *left hand* side gives the imputed values from one set of imputations, and the *right hand* side gives the complete data

Method 1



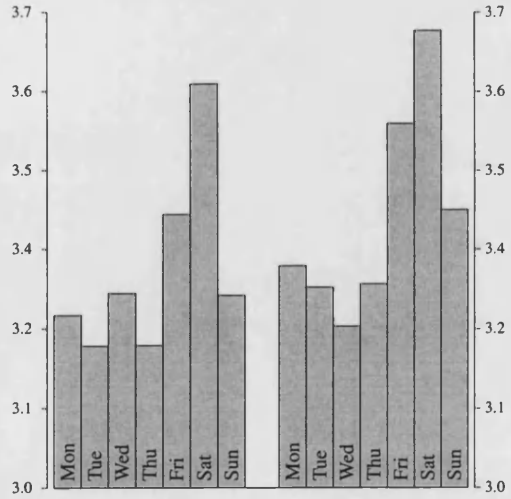
Imputed *Complete Data*

Method 2



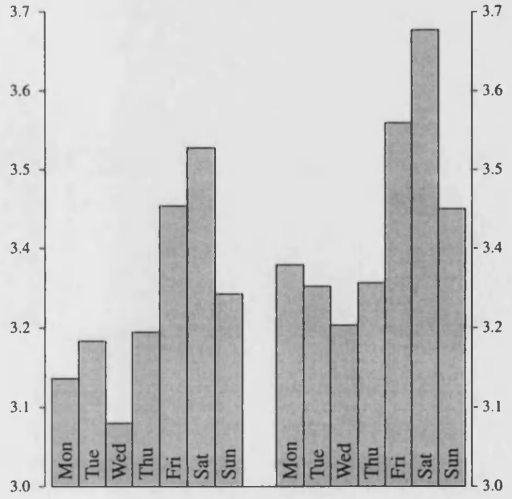
Imputed *Complete Data*

Method 3



Imputed *Complete Data*

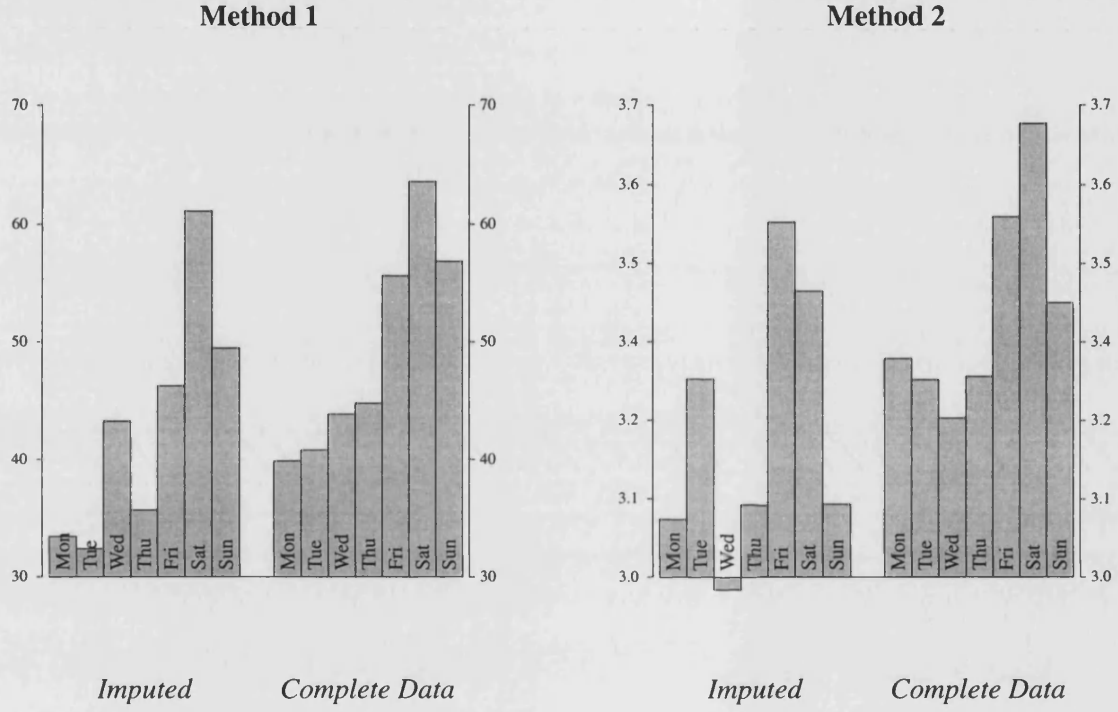
Method 4



Imputed *Complete Data*

Figure 6.6: Patterns of signs (proportions of men drinking on each day of the week) and of amounts (mean of log-transformed positive amounts drunk by men on each day of the week) for Method 5 (MIX with DA, using a restricted independence model)

Each graph consists of two parts: the *left hand* side gives the imputed values from one set of imputations, and the *right hand* side gives the complete data



6.7.4 Discussion

Using Schafer's software, the excessive bias encountered with the SOLAS procedures, due to the need to use theoretically unsuitable methods, is avoided. All the imputation methods tested using Schafer's software produced only small biases in the estimates of alcohol consumption using MCAR and MAR datasets. The pattern of consumption was not preserved by Method 1, and that of Method 3 was an improvement on Method 2. These results indicate that more information about pattern is preserved by ordering the diary by day of the week. As expected, the Methods that ignore the start day of the diary (Methods 3 and 4) did not seem to present any disadvantage in terms of the bias of the estimates used here, or of the preservation of drinking pattern over the week. Method 5 failed to reflect the pattern of drinking, not surprisingly as it used a restricted independence model.

It was conjectured that MIX would be the best procedure to use, as this software treats categorical and continuous variables appropriately, and can use the joint distribution of signs and amounts to impute the missing values in a single step. Method 4, using MIX, gives lower bias than the other methods and also appears to preserve the pattern of drinking over the week. Any gains in using MIX were, however, outweighed by the increasing complexity of the model. For 'proper' imputations (Section 2.7.2; Rubin, 1987; Schafer, 1997) the uncertainty in the estimated parameter values should be reflected in the imputations, and using Schafer's procedures this requires the Data Augmentation (DA) step. The DA step could only be included with MIX when the number of categorical variables was restricted, and an independence model used (Method 5). However, this Method was biased under MCAR and did not reflect the pattern of drinking over the week.

Dividing the imputation process into two steps using CAT and NORM with an unrestricted model (Method 3) provides the best compromise. With this Method it is possible to use DA without simplifying the model. The results of applying Method 3 with the DA step are presented in Table 6.17 below. Comparing these results with those for Method 3 without DA in Tables 6.15 and 6.16, the standard errors are somewhat larger, reflecting the additional uncertainty in the parameter estimates, whilst the biases remain small compared to the other methods. Method 3 with the addition of DA was therefore chosen to apply to the NSHD data.

6.8 Sensitivity to the MAR assumption

6.8.1 Introduction

We now look at what the results of using the simulated MNAR (missing not at random) data can tell us. All the methods of imputation depend on the assumption that the missing values are missing at random (MAR) given the information in the covariates and the observed values of the variable to be imputed. We have seen that the MAR assumption is the key to exploiting the information about the missing values contained in the incomplete records. We would not expect any method to yield unbiased estimates when applied to MNAR data, since they all assume that

Table 6.17: Method 3 with DA: estimates of proportions over weekly and daily limits for MCAR and MAR data

Estimated proportions with alcohol consumption over weekly limits

	Complete data	MCAR data					MAR data				
Women	**%**	**BIAS**	**RMSE**	**SE**	**95% CI**	***F***	**BIAS**	**RMSE**	**SE**	**95% CI**	***F***
>14	**15.6**	−0.14	0.54	0.52	(−1.20,0.92)	0.07	−0.55	0.87	0.67	(-1.92,0.82)	0.67
>35	**1.1**	0.08	0.21	0.20	(−0.33,0.49)	0.16	0.04	0.20	0.20	(−0.37,0.45)	0.04
Men											
>21	**33.2**	−0.66	0.97	0.71	(−2.10,0.78)	0.87	−0.96	1.26	0.81	(−2.62,0.70)	1.40
>50	**8.6**	−0.11	0.45	0.44	(−1.01,0.79)	0.06	−0.52	0.71	0.49	(−1.52,0.48)	1.13

Estimated proportions with alcohol consumption over daily limits

	Complete data	MCAR data					MAR data				
Women	**%**	**BIAS**	**RMSE**	**SE**	**95% CI**	***F***	**BIAS**	**RMSE**	**SE**	**95% CI**	***F***
>3	**38.1**	0.13	0.59	0.57	(−1.04,1.31)	0.05	−0.02	0.65	0.65	(−1.35,1.30)	0.00
>6	**11.5**	0.98	1.08	0.44	**(0.08,1.89)**	4.93	1.62	1.73	0.60	**(0.39,2.86)**	7.22
Men											
>4	**63.3**	−0.96	1.22	0.76	(−2.51,0.60)	1.58	0.81	1.21	0.90	(−1.04,2.66)	0.81
>8	**35.0**	−0.90	1.16	0.74	(−2.41,0.62)	1.47	−0.69	0.99	0.71	(−2.14,0.76)	0.94

Percentage points of $F_{1,29}$: $P = 0.05$ $F = 4.18$, $P = 0.01$ $F = 7.60$, $P = 0.001$ $F = 13.4$

the data is MAR. However, the MAR assumption cannot be tested empirically, so we cannot establish whether it holds or not. MNAR cannot be ruled out, and with some data, such as alcohol consumption, it seems entirely plausible that survey subjects might not be able to complete a diet diary simply because they had had too much to drink.

It has been argued that multiple imputation can protect against the consequences of MNAR (Section 1.3.4). We can evaluate this contention only using simulated MNAR data .

6.8.2 Methods

Three of the methods studied in this chapter were applied to the MNAR datasets. The methods are listwise deletion (LD) (Section 6.3.2), and two methods of multiple imputation: the SOLAS model based procedures (Section 6.6.2) and Method 3 using Schafer's CAT and NORM procedures (Section 6.7.2). These three methods are applied to different numbers of repetitions of the simulation process (Section 6.2.1): $n = 100$ for LD, $n = 11$ for SOLAS and $n = 30$ for Schafer Method 3. Recall that the MNAR simulation items of alcohol consumption in the diary

were deleted with a probability proportional to reported alcohol consumption on the days deleted (Section 6.2.3).

6.8.3 Results

The results for the three methods are presented in Table 6.18.

We would expect a negative bias in the results, because the records for the diary days when people drank heavily are more likely to have been deleted. This is shown in the LD results. All biases are negative and substantial, with the upper limit of the confidence interval below zero. For SOLAS model based procedures, the estimated biases are much smaller, and generally positive. Even though these results are based on very few replicates, each of the replicates produced an estimate in the same direction except for just two of the 88 estimates, i.e. 11 replicates of 8 estimates. The 95% confidence intervals for three of the proportions indicate that the bias could have arisen by chance (they include zero). In these estimates the bias is still greater than one percent and they are based on only 11 repetitions of the process so we cannot be very confident in the estimates. Using Schafer's procedures the biases are negative but substantially smaller than for LD, though in general not as small as those for SOLAS. Some of these biases are very small, under one percent, and the confidence intervals imply that these could have arisen by chance.

6.8.4 Discussion

It is tempting to conclude that the results for SOLAS procedures indicate that this method would protect against the possibility of MNAR, as suggested by Rubin (1996). However, we have found that the results for the SOLAS procedures are positively biased when applied to MCAR data (Section 6.6.3) and that this arises because of a problem with the algorithm used (Section 6.6.4). This explains the lack of negative bias in the MNAR results. The positive bias in the method is simply reduced compared to that using MCAR or MAR data (Table 6.6).

However, the Schafer method is known to be unbiased under MCAR (Table 6.15) and MAR (Table 6.16). The reduction in negative bias compared to those for listwise deletion indicates that the method protects against the consequences of MNAR. The method has succeeded in capturing the information in the incomplete records to 'recover' levels of drinking. Even though subjects who drank more on the days when they did not 'report' their alcohol consumption, we have 'found out' some of these heavier drinkers because of the information they provided, perhaps in their reported weekly recall or their drinking in the first two diary days. As suggested (Schafer, 1997), these methods for multiple imputation, by exploiting information in the incomplete records, perform better than naïve ones even when the data is MNAR.

6.9 The method of choice

This chapter has investigated various ways of dealing with item non-response to alcohol consumption in a diet diary. The motive was to choose a method to deal with this problem in practice using the NHSD data for the 3262 respondents interviewed in 1989. Here the method

Table 6.18: Listwise deletion and and SOLAS Model-based Methods, and Schafer Method 3, estimates of proportions over weekly and daily limits from MNAR data: comparison with proportions estimated from complete data

Estimated proportions with alcohol consumption over weekly limits: MNAR data

	Complete data	LD				SOLAS Model Based				Schafer METHOD 3			
Women	**%**	**BIAS**	**RMSE**	**SE**	**95% CI**	**BIAS**	**RMSE**	**SE**	**95% CI**	**BIAS**	**RMSE**	**SE**	**95% CI**
>14	**15.6**	−7.88	7.92	0.74	(−9.35, −6.41)	1.34	1.58	0.85	(−0.55, 3.23)	−2.25	2.31	0.55	(−3.37,−1.13)
>35	**1.1**	−0.77	0.80	0.23	(−1.23, −0.31)	1.17	1.22	0.34	(0.41, 1.93)	0.29	0.36	0.21	(−0.14, 0.72)
Men													
>21	**33.2**	−9.93	10.03	1.38	(−12.67, −7.19)	2.80	2.99	1.07	(0.42, 5.18)	−2.95	3.02	0.68	(−4.34,−1.56)
>50	**8.6**	−3.35	3.42	0.69	(−4.72, −1.98)	1.42	1.74	1.01	(−0.83, 3.67)	−0.67	0.78	0.38	(−1.45, 0.11)

Estimated proportions with alcohol consumption over daily limits: MNAR data

	Complete data	LD				SOLAS Model Based				Schafer METHOD 3			
Women	**%**	**BIAS**	**RMSE**	**SE**	**95% CI**	**BIAS**	**RMSE**	**SE**	**95% CI**	**BIAS**	**RMSE**	**SE**	**95% CI**
>3	**38.1**	−15.01	15.06	1.16	(−17.31,−12.71)	−2.80	3.03	1.16	(−5.38,−0.22)	−5.25	5.28	0.61	(−6.50,−4.00)
>6	**11.5**	−5.61	5.65	0.61	(−6.82, −4.40)	2.52	2.63	0.75	(0.85, 4.19)	−0.27	0.56	0.49	(−1.27, 0.73)
Men													
>4	**63.3**	−12.92	13.00	1.44	(−15.78,−10.06)	−1.77	1.87	0.60	(−3.11,−0.43)	−5.09	5.14	0.68	(−6.48,−3.70)
>8	**35.0**	−10.36	10.44	1.24	(−12.82, −7.90)	1.21	1.56	0.98	(−0.97, 3.39)	−3.24	3.31	0.67	(−4.61,−1.87)

chosen is described. The advantages of this method, relative to the others investigated in this chapter, have already been discussed in the preceding sections of this chapter. Here we describe the method of choice.

The method uses Schafer's procedures for multiple imputation of the missing values in the daily items of alcohol consumption in the diet diary. The seven items of alcohol consumption in the diary are first ordered by the day of the week. The method uses a two-step process which first imputes the sign of alcohol consumption using Schafer's CAT and then the positive amount of alcohol consumption using NORM. The output from each procedure is a set of 5 completed datasets. The values for the log-transformed positive amounts are first transformed back by exponentiating. The signs (from CAT) and amounts (from NORM) are then combined by multiplying the sign by the corresponding amount from each pair of datasets in turn (Section 6.5.2), to give five completed datasets for the alcohol consumption of each diary day.

The variables used in the CAT procedure to impute the sign of alcohol consumption are: sign of alcohol consumption on the available seven diary days, gender, smoking status (smoker or non-smoker), sign of weekly recall (whether the subject consumed any alcohol in the previous week or not), adult social class (manual or non-manual), and CAGE (coded in the two categories 0–1 and 2–4). The variables used in the NORM procedure for the imputation of positive quantities are log-transformed positive amounts of alcohol consumption (in grams) on the available seven diary days, gender, smoking status, weekly recall amount (in units, log-transformed after adding unity), CAGE score, and adult social class.

In each procedure the following steps are executed: a preliminary computation step, an EM step, a Data Augmentation (DA) step and an imputation step. Five sets of imputed values are obtained by replicating the DA and imputation step. For CAT, the saturated model is used, with a limit of 1000 iterations. The R versions of these programs were used. These are available from http://www.stats.bris.ac.uk/R/ . The method could be implemented using S-Plus in exactly the same way. The S-Plus library 'missing' has the procedures 'Loglin' and 'Gauss' which implement Schafer's algorithms for CAT and NORM respectively.

The five completed datasets can each be analysed by complete data methods of analysis, using standard software. One simply repeats the analysis on each of the datasets giving five estimates, and combines the resulting estimates using the methods given in Section 2.9, as demonstrated in Section 6.5.3.

The sensitivity of this method to whether missing data are MAR has been assessed in Section 6.8. The results suggest that the method mitigates against the consequences of MNAR. In the next chapter, we examine the sensitivity of inferences about alcohol consumption to dealing with missing data using this method. By varying the model for imputation (the variables included in the imputation procedures), the sensitivity of inferences to the imputation model is assessed.

Chapter 7

Analysis of Alcohol Consumption in the MRC National Survey of Health and Development

7.1 Introduction

This chapter reports the results of applying the method developed and evaluated in Chapter 6 to the NSHD data collected in 1989. The object is to make inferences to the population well represented by the 3262 subjects interviewed in 1989 at age 43. We interpret this population broadly as all native-born residents of the UK, born shortly after the end of World War II.

We take account of two aspects of missing data on alcohol consumption in the diary. First, some subjects did not complete all the days in the diet diary — this is referred to as item non-response. Second, case non-response arises because some of the members of the original birth cohort were not included, by design, in the sample (only 1 in 4 of births to wives of manual workers were included, as opposed to all births to wives of non-manual and agricultural workers). This is referred to as missing by design (Section 2.5.2), and since the mechanism of the 'missingness' is known, it can be dealt with by standard methods for survey analysis (Section 2.6.1).

Item non-response is dealt with by multiple imputation (MI), using the method developed in Chapter 6. We complete the records of alcohol consumption on each day of the seven-day diary by imputation. The alcohol consumption is modelled by exploiting the dependence of the recorded alcohol consumption on observed covariates. To represent the uncertainty in the missing data, multiple sets of plausible values are generated. These plausible values are drawn independently from the distribution implied by the model for alcohol consumption.

In this chapter we use multiply imputed datasets to estimate the prevalence of excessive alcohol consumption in mid-life for the post-war population, and to estimate the dependence of systolic blood pressure on birthweight in men (Chapter 4), and the association of alcohol consumption and blood pressure.

In Section 6.7, the method was shown to yield unbiased estimates of excessive alcohol consumption when the data is MCAR or MAR (conditional on weekly recall), and to preserve the pattern of alcohol consumption over the days of the week. In Section 6.8 it was shown that, even if the data is MNAR, the estimates of excessive alcohol consumption are relatively unbiased compared with using complete records only. However, it is unlikely, in practice, that a

model for imputation could be assessed by simulation, and it is important to acknowledge our uncertainty about the model for imputation by performing sensitivity analysis.

The value of the imputation of alcohol consumption in this thesis is that the multiply imputed datasets can be used in further epidemiological analyses involving alcohol consumption. However, the validity of multiple imputation inferences have been questioned when the analysis includes a variable that was not included in the model for imputation (Section 1.3.5). Also, as MI becomes available in standard software packages, and can be more widely utilised, it may become the practice to impute at the point of analysis. This may lead analysts to impute using only the variables in their model for analysis. In this chapter we assess the sensitivity of the estimates to different models for imputation: excluding a variable used in the analysis and using only the variables included in the analysis.

7.2 Methods

The first model for multiple imputation is specified in Section 6.9. The covariates for this model are: gender, observed alcohol consumption on the diary days, alcohol consumption reported in the weekly recall, adult social class, CAGE score, smoking status, and the day of the week by virtue of ordering the diary records in weekday order. In the second model, systolic blood pressure (SBP) is included in addition. Thus, the first model is referred to as the imputation model '*without SBP*', the second as that '*including SBP*'.

Another imputation model used is referred to as the '*analyst's model*', and this includes only the variables to be used in the model for the analysis. For example, in the regression model for birthweight on systolic blood pressure (Chapter 4) the variables used are those specified in Section 4.2: SBP, birthweight, childhood social class, current social status, BMI, exercise and level of alcohol consumption. This imputation uses only the NORM procedure. Levels of alcohol consumption imputed are rounded to the nearest category.

Another example used is the (unadjusted) association between level of alcohol consumption and systolic blood pressure (SBP). In this example, the 'analyst's model' comprises only the variables SBP and observed total alcohol consumption in the diary week. The imputation model is applied separately for men and women since the relationship between alcohol consumption and SBP differs according to gender (Section 4.3.1). The model is implemented using MIX with the sign of total alcohol consumption as a categorical variable and logged positive amount of total alcohol consumption (with zeros set to missing) and SBP as continuous variables. The positive amounts are exponentiated and multiplied by the sign. Note that this model imputes only a total alcohol consumption for the week, without generating values for each day. This model is referred to as the imputation model 'SBP only'.

All the imputation methods use Data Augmentation (Section 6.7.2). Five imputations are used. The efficiency of an estimate increases with the number of imputations (m). For maximum

efficiency of estimation an infinite number of imputations would be required. The relative efficiency of using only m imputations depends on the rate of missing information for the quantity being estimated, as well as m. Hence the justification for choosing $m = 5$ is post hoc, and is given in Section 7.3.1. In general, quite small numbers of imputations suffice to give relatively high efficiency. For example, even when 50% of information is missing, $m = 5$ imputations gives an approximate relative efficiency of 91% (Equation (8) in Section 2.9).

Each set of imputations completes the data for the seven-day diary records as well as any missing values of the covariates used in the imputation model. The advantage of MI-estimation is that each of the m completed datasets is analysed using the same standard complete data method. The m results are then combined by a simple operation. Here, the five completed datasets are analysed in SPSS. To compensate for the unequal representation arising from the stratification of the birth sample, weighting is applied. The required estimate and its variance are derived for each stratum separately. The estimates for each stratum are then combined using a weighted combination of the estimates for the two strata as described in Section 2.6.1. The analysis is repeated on each of the five completed datasets. The five estimates and their standard errors are combined to give an MI-estimate using the formulae given in Section 2.9.

We have already referred to the way in which categories of total drinking in the week are defined in Table 4.1 (Section 4.2). The estimates of interest are of the proportions of the population drinking above the recommended limits, defined in Section 6.2.4.

The relationship of birthweight and systolic blood pressure (SBP) in men is given in terms of the coefficient for birthweight in the regression model with SBP as the dependent variable, controlling for childhood social class, current social status, BMI, exercise and alcohol consumption, as described in Section 4.2. The estimates of the contrasts of mean blood pressure between different levels of alcohol consumption are obtained from regression in SPSS. Using the level of total alcohol consumption in the diary week (None, Sensible, Immoderate, Heavy) as a categorical covariate, SBP was regressed on the level of alcohol consumption. The contrast between two levels is the difference of the coefficients for these levels, and the standard error of this contrast is obtained from the variances and covariance of these coefficients.

7.3 Results

7.3.1 Prevalence of excessive alcohol consumption

Using multiple imputation to deal with item non-response

In order to explore the properties of the MI-estimates, the unweighted estimates based on MI are compared with those based on complete records (listwise deletion, LD). Table 7.1 gives the estimates of the proportions drinking over the limits (specified in Section 7.2) based on MI, using the imputation model without SBP, averaged over the five datasets (equation (1) in Section 2.9). The MI-estimates are all somewhat higher than those for LD; for example, the MI-estimate for women drinking heavily on any day is 13.0% compared with 11.5% for LD. This is

Table 7.1: Comparison of estimates of proportions consuming in excess of weekly and daily limits: use of complete records (LD) versus multiple imputation (MI)

	Listwise Deletion (LD)				Multiple Imputation (MI)			
	n	%	se	95% CI	*n*	%	se	95% CI
Estimates of proportions exceeding weekly limits								
Women	1024				1627			
Excessive		15.6	1.13	(13.4,17.8)		16.0	0.98	(14.1,17.9)
Heavy		1.1	0.32	(0.4, 1.7)		1.1	0.30	(0.5, 1.7)
Men	978				1635			
Excessive		33.2	1.51	(30.3,36.2)		35.0	1.32	(32.5,37.6)
Heavy		8.6	0.90	(6.8,10.3)		9.4	0.79	(7.9,11.0)
All	2002				3262			
Estimates of proportions exceeding daily limits								
Women	1024				1627			
Excessive		38.1	1.52	(35.1,41.1)		39.2	1.32	(36.6,41.7)
Heavy		11.5	1.00	(9.6,13.5)		13.0	1.09	(10.8,15.1)
Men	978				1635			
Excessive		63.3	1.54	(60.3,66.3)		64.8	1.22	(62.5,67.2)
Heavy		35.0	1.52	(32.0,38.0)		37.2	1.31	(34.7,39.8)
All	2002				3262			

Note: In the tables in this chapter, levels of alcohol consumption are indicated as follows:

Exceeding weekly limits:

Excessive drinking

Women	**Men**	
>14 U		More than 14 units consumed in total during the week
	>21 U	More than 21 units consumed in total during the week

Heavy drinking

Women	**Men**	
>35 U		More than 35 units consumed in total during the week
	>50 U	More than 50 units consumed in total during the week

Exceeding daily limits:

Excessive drinking

Women	**Men**	
>3 U		More than 3 units consumed in a day, on at least 1 day
	>4 U	More than 4 units consumed in a day, on at least 1 day

Heavy drinking

Women	**Men**	
>6 U		More than 6 units consumed in a day, on at least 1 day
	>8 U	More than 8 units consumed in a day, on at least 1 day

NB: 'Excessive drinking' includes 'Heavy drinking'

not unexpected, since it has been observed that factors associated with higher alcohol consumption were also associated with having an incomplete diary (Section 5.4).

The MI-estimates are, in general, more efficient than those using LD because they draw on more information. Apart from the estimated proportion of women drinking heavily on any day, the standard errors for the MI-estimates are smaller than those for LD. This is a result of MI exploiting the information in the incomplete records. The proportional reduction in the standard errors varies between the estimates: from 21% (1 – 1.22/1.54) for the estimated proportion of men drinking excessively on any day to 6% (1 – 0.30/0.32) for that of women drinking heavily during the week. The standard error for the estimated proportion of women drinking heavily on any day is inflated by 9%. The reduction of the standard errors of the MI-estimates relative to LD indicates the contribution to the estimation that can be attributed to the incomplete records.

The components of the variance of the MI estimates are detailed in Table 7.2. They are formally defined by Rubin (1987) and are given in Section 2.9.

Table 7.2: Components of variance of MI estimates of proportions exceeding weekly and daily limits

Estimate	$\bar{U}$	B	T	r	γ (%)	RE
Estimates of proportions exceeding weekly limits						
Women						
Excessive	0.82	0.11	0.96	0.16	14.6	0.97
Heavy	0.07	0.02	0.09	0.28	23.9	0.95
Men						
Excessive	1.39	0.30	1.75	0.25	21.9	0.96
Heavy	0.52	0.09	0.63	0.21	18.6	0.96
Estimates of proportions exceeding daily limits						
Women						
Excessive	1.46	0.23	1.74	0.19	16.9	0.97
Heavy	0.69	0.42	1.19	0.72	46.3	0.92
Men						
Excessive	1.39	0.08	1.49	0.07	6.6	0.99
Heavy	1.43	0.25	1.73	0.21	18.4	0.96

The MI-estimate of the sampling variance is given by the 'within' variance of the five imputation estimates, denoted by $\bar{U}$ in equation (2) in Section 2.9. This is the variance that would be obtained had the data been complete for all 3262 respondents. The estimation of the sampling variance is more efficient using MI ($n = 3262$) than LD ($n = 2002$) because more cases are used. However, the uncertainty is greater than if the data were complete because some

records are incomplete. The additional uncertainty due to nonresponse is measured by the variance of the five imputation estimates, the between-imputation variance denoted by B in equation (3) in Section 2.9. The total variance (T) is the sum of these two components, with an adjustment of B to take into account the number of imputations (equation (4) in Section 2.9). The MI-estimate of the sampling variance, T, takes account of the uncertainty due to missing data. The MI-estimate standard errors, given in Table 7.1, are calculated as $\sqrt{T}$, and hence take account of the uncertainty due to missing data in the estimate.

The between-imputation variance component (B) is generally small relative to the estimated sampling variance ($\bar{U}$). So the additional uncertainty due to nonresponse is small relative to our uncertainty about the population value based on the sample estimate. The comparison of these components is given by the relative increase in the variance due to non-response, denoted by r in equation (5) in Section 2.9. This varies with the estimate: for example, there is relatively more uncertainty due to non-response about the proportion of women drinking heavily on any day ($r = 0.72$) and over the week ($r = 0.28$) than for other estimates, whilst this is low for men drinking excessively on any day ($r = 0.07$). Gamma (γ), a function of r, essentially compares the between-imputation variance component to the total variance (equation (6) in Section 2.9). This is called the fraction of missing information and is given as a percentage in Table 7.2. The percentage of missing information is generally much less than the proportion of people with incomplete diaries (40.2% of men and 37.1% of women, Table 5.1). This is because the MI method has utilised the information in the partially complete diaries and in the covariates that are used as auxiliary information in the imputation model. Note that the percentage of missing information about each estimate is quite different for the different estimates: from as little as 6.6% about men drinking excessively over the week to 46.3% about women drinking heavily on any day. The observed data in the diary and in the covariates used in the imputation model are not as informative about some estimates as about others. One would expect the proportion of missing information to be higher for women drinking heavily, because there are relatively few women heavy drinkers in the observed data, and a higher proportion of them have incomplete records than women in general.

The relative efficiency (RE) gives the approximate efficiency of using the five-imputation estimator compared with using an infinite number of imputations (equation (8) in Section 2.9). These results, all above 92%, show that using five imputations is sufficient and there would be relatively small advantage in increasing the number of imputations.

Using weighting to deal with the stratified sample design

Table 7.3 gives the level of alcohol consumption in the two sample strata for the 2002 men and women who completed their dairy. Men who at birth had fathers in manual (but not agricultural) occupations are, on average, less likely to drink sensibly (48.1%), than those whose fathers had been non-manual or agricultural workers (54.6%). The difference between the strata is greater for heavy drinking (11.1% for those with non-manual fathers compared with 6.9% for

Table 7.3: Comparison of levels of alcohol consumption between sample strata in completers

Level of total alcohol consumption in the diary week		Sample stratum: Non-manual and agricultural		Sample stratum: Manual (but not agricultural)	
		n	%	n	%
Women	**None**	166	27.4	122	29.2
	Sensible	344	56.8	232	55.5
	Immoderate	86	14.2	63	15.1
	Heavy	10	1.7	1	0.2
	Total	606	100.0	418	100.0
Men	**None**	87	15.0	58	14.6
	Sensible	317	54.6	191	48.1
	Immoderate	137	23.6	104	26.2
	Heavy	40	6.9	44	11.1
	Total	581	100.0	397	100.0

those with manual fathers) than for immoderate drinking (26.2% vs. 23.6%). For women the differences between the strata, although on average smaller, tend to be in the opposite direction to those for men. Women from a higher social class at birth tended to drink more sensibly, rather than abstain during the diary week (56.8% and 27.4% compared with 55.5% and 29.2%), and heavily rather than immoderately (1.7% and 14.2% compared with 0.2% and 15.1%). Since alcohol consumption varies between the sample strata of the birth cohort, weighting to compensate for the unequal representation of the strata affects estimates for the population as a whole. The weighted proportion from each dataset is combined to give the MI-estimate of weighted results (on the right-hand side of Table 7.4). The estimated proportions of men drinking excessively either during the week or on any day are greater for the weighted analysis (36.6% during the week, 66.7% on any day) than for the unweighted analysis (35.0% during the week, 64.8% on any day). Similarly, estimates are greater using weighted compared with unweighted analysis for men drinking heavily. For women, the estimates differ little, but the differences are mainly in heavy drinking (Table 7.3) and the number of women drinking heavily is small. In the following sections, all analyses are weighted to give estimates for the population represented by the respondents interviewed in the 1989 survey.

Table 7.4: Comparison of estimates of proportions drinking in excess of weekly and daily limits: MI-estimates from unweighted and from weighted analyses

	MI-estimates unweighted				MI-estimates weighted			
	n	%	se	95% CI	*n*	%	se	95% CI
Estimates of proportions exceeding weekly limits								
Women	1627				1627			
Excessive		16.0	0.98	(14.1,17.9)		15.9	1.20	(13.5,18.2)
Heavy		1.1	0.30	(0.5, 1.7)		1.0	0.38	(0.2, 1.7)
Men	1635				1635			
Excessive		35.0	1.32	(32.5,37.6)		36.6	1.54	(33.5,39.6)
Heavy		9.4	0.79	(7.9,11.0)		10.4	0.96	(8.5,12.3)
All	3262				3262			
Estimates of proportions exceeding daily limits								
Women	1627				1627			
Excessive		39.2	1.32	(36.6,41.7)		39.3	1.62	(36.1,42.5)
Heavy		13.0	1.09	(10.8,15.1)		13.0	1.39	(10.3,15.7)
Men	1635				1635			
Excessive		64.8	1.22	(62.5,67.2)		66.7	1.43	(63.8,69.5)
Heavy		37.2	1.31	(34.7,39.8)		39.6	1.49	(36.7,42.5)
All	3262				3262			

7.3.2 The relationship between birthweight and systolic blood pressure in mid-life

We reported in Chapter 4 that the exclusion of subjects who did not complete their diary had a selection effect. It was estimated that a 1 kg increase in birthweight is associated with a 3.79 mm Hg lower average systolic blood pressure (95% confidence interval 1.45 to 6.13), based on the 912 cases with complete records. Table 7.5 compares this estimate with that of the MI-estimate based on the imputed datasets from the imputation model without SBP. The corresponding MI-estimate, including all subjects who completed the diary, is 1.70 mm Hg. The 95% confidence interval, −3.48 to 0.08 (P = 0.06), indicates that we can be fairly confident that there is a negative association between birthweight and blood pressure, but the degree of association is much weaker than would be inferred using only cases with complete records.

Table 7.5: Regression coefficient for birthweight on systolic blood pressure for men: comparison of complete records (LD) and multiple imputation (MI)

	Regression coefficient for birthweight				
	n	β	se	*P*	95% CI
Complete cases (LD)	912	−3.79	1.19	0.002	(−6.13,−1.45)
MI:					
imputation model without SBP	1499	−1.70	0.92	0.061	(−3.48, 0.08)

Linear Model for systolic blood pressure in men: coefficient β for birthweight (in kg), controlling for: childhood social class, current social status, BMI, exercise and alcohol consumption

We now assess the sensitivity of the MI-estimate to the imputation model used, by comparing three imputation models. The first imputation model does not include SBP (*'without SBP'*), although the association between alcohol consumption and SBP motivated its inclusion in the regression model. The second imputation model uses the same method as the first but includes the variable SBP in addition to the other covariates (*'including SBP'*). The third imputation model is the *'analyst's model'* in which only the variables used in the regression model are included in the imputation model (Section 7.2). The MI-estimates using these three imputation models are given in Table 7.6. Using the imputation model 'including SBP' yields an MI-estimate of 1.62 mm Hg lower average SBP for a 1 kg increase in birthweight (95% CI −3.41 to 0.16), compared with the estimate for the 'analyst's model' of 1.72 mm Hg (95% CI −3.44 to 0.01), and 1.70 mm Hg for the imputation model 'without SBP'. The similarity of these results indicates the lack of sensitivity (that is, robustness) of the estimated regression coefficient to the imputation model. This is hardly surprising since there was only a small proportion of missing data for the variables SBP and birthweight, and the indication (based on the completers) was that the association between these variables did not depend on the level of alcohol consumption (Section 4.3.5).

7.3.3 The association between alcohol consumption and systolic blood pressure

For men who completed their diaries, SBP was found to increase monotonically with the level of total alcohol consumption in the diary week, while for women there was a suggestion of a 'U' shaped relationship (Section 4.3.1). We now examine, for men and women separately, the difference in mean SBP between levels of alcohol consumption, contrasting sensible drinkers with those who drank nothing, immoderate drinkers with sensible drinkers, and heavy drinkers with immoderate drinkers.

Results for men

Data on both alcohol consumption and blood pressure are recorded for 961 men. For these men, SBP is on average 1.79 mm Hg higher for sensible drinkers than for those who drank nothing, 2.12 mm Hg higher for immoderate than for sensible drinkers and 4.15 mm Hg higher for heavy drinkers than for immoderate drinkers (Table 7.7, LD). Mean SBP seems to increase with the

Table 7.6: Regression coefficient for birthweight on systolic blood pressure for men, using multiple imputation: sensitivity to the imputation model

	Regression coefficient for birthweight				
	n	β	se	*P*	95% CI
MI:					
Imputation model without SBP	1499	−1.70	0.92	0.061	(−3.48, 0.08)
MI:					
Imputation model including SBP	1529	−1.62	0.91	0.074	(−3.41, 0.16)
MI:					
Imputation model=Analyst's model	1635	−1.72	0.88	0.051	(−3.44, 0.01)

Linear model for systolic blood pressure for men: coefficient β for birthweight (in kg), controlling for: childhood social class, current social status, BMI, exercise and alcohol consumption

level of alcohol consumption, and the increase is particularly large for heavy drinkers in contrast with moderate drinkers. However, the relatively large standard errors and wide confidence intervals indicate that there is considerable uncertainty in these estimates. With the exception of heavy drinkers, there is some doubt that blood pressure increases with alcohol consumption.

We now compare the estimates based on MI and LD (complete records) of differences in mean SBP (Table 7.7). The MI-estimates are based on the imputed values for alcohol consumption using the first imputation model, which does not include SBP. The MI-estimates indicate that there are larger differences, compared with LD estimates, in mean blood pressure between sensible drinkers and those who did not drink (2.24 mm Hg, MI 25% larger than LD), and for immoderate drinkers compared with sensible drinkers (2.29 mm Hg, MI 8% larger than LD), but the difference for heavy compared with immoderate drinkers is lower (3.11, MI 8% lower than LD). The MI-estimates are more efficient and the lower estimated standard errors give us more confidence in the increases in mean blood pressure, although the 95% confidence intervals indicate that the possibility of a very small decrease in blood pressure with increasing alcohol consumption level cannot be ruled out.

Table 7.7: Increase in SBP between successive levels of alcohol consumption for men: comparison of complete cases (LD) and MI

	LD ($n = 961$)				MI (without SBP) ($n = 1587$)			
Contrast	**estimate**	**se**	**95% CI**	***P***	**estimate**	**se**	**95% CI**	***P***
Sensible *vs* none	1.79	1.77	(−1.68,5.25)	0.313	2.24	1.55	(−0.80,5.28)	0.149
Immoderate *vs* sensible	2.12	1.44	(−0.69,4.94)	0.140	2.29	1.39	(−0.44,5.01)	0.100
Heavy *vs* immoderate	4.15	2.17	(−0.10,8.41)	0.056	3.11	1.66	(−0.15,6.37)	0.062

The problem with the MI-estimate using an imputation model without SBP is that the imputation model has not taken into account the relationship between alcohol consumption and SBP. We now impute alcohol with SBP included in the model, to assess the sensitivity of these results to the imputation model. The results for the two imputation models are given in Table 7.8.

Table 7.8: Increase in SBP between successive levels of alcohol consumption for men: sensitivity to the model used for multiple imputation

	Multiple Imputation Model							
	Without SBP ($n = 1587$)				Including SBP ($n = 1635$)			
Contrast	**estimate**	**se**	**95% CI**	***P***	**estimate**	**se**	**95% CI**	***P***
Sensible *vs* none	2.24	1.55	(−0.80,5.28)	0.149	2.46	1.43	(−0.35,5.27)	0.086
Immoderate *vs* sensible	2.29	1.39	(−0.44,5.01)	0.100	2.83	1.15	(0.58,5.08)	0.014
Heavy *vs* immoderate	3.11	1.66	(−0.15,6.37)	0.062	3.56	1.94	(−0.25,7.37)	0.067

The MI-estimates 'including SBP' impute SBP in addition to alcohol consumption, so that all the 1635 men interviewed in 1989 are included in the analysis. Compared with the results using the imputation model 'including SBP', those without SBP are biased towards zero. For example, when SBP is included in the imputation model, the MI-estimate of the mean difference in SBP for immoderate compared with sensible drinkers is 2.83 mm Hg, compared with 2.29 mm Hg when SBP is not included in the imputation model. This increase in mean SBP is estimated with more precision, so we have a high degree of confidence ($P = 0.014$) in the increase (95% CI 0.58 to 5.08). The MI-estimates 'including SBP' indicate that increasing levels of alcohol consumption are associated with increases in mean blood pressure, and that these increases are greater for higher levels of alcohol consumption. Mean SBP for men

increases by 2.46, 2.83 and 3.56 mm Hg as levels of drinking increase from none to sensible, immoderate and heavy, respectively.

Results for Women

Data on both alcohol consumption and blood pressure are recorded for 998 women. For these women, SBP is on average 0.37 mm Hg lower for sensible drinkers than for those who drank nothing, 3.47 mm Hg higher for immoderate than for sensible drinkers and.26.81 mm Hg higher for heavy drinkers than for immoderate drinkers (Table 7.9, LD). There is no evidence that the relationship is 'U' shaped in this weighted estimate. The lower mean SBP for sensible drinkers compared to those who did not drink occurs only in the sample stratum of women with fathers in non-manual (or agricultural) occupations (results not presented). There is stronger evidence of an increase in mean SBP for excessive drinkers — above the recomended 14 Units per week ($P = 0.06$ for immoderate vs. sensible and $P = 0.027$ for heavy vs. immoderate). There are very few women with complete data who drank heavily ($n = 9$: 8 with non-manual fathers and only 1 with a father in a manual occupation), and the very wide confidence interval indicates that the very large estimated difference in mean SBP for heavy vs. immoderate drinkers (26.81 mm Hg) is not reliably estimated.

Table 7.9: Increase in SBP between successive levels of alcohol consumption for women: comparison of complete cases (LD) and MI

	LD (n=998)				MI (without SBP) (n=1570)			
Contrast	**estimate**	**se**	**95% CI**	***P***	**estimate**	**se**	**95% CI**	***P***
Sensible *vs* none	−0.37	1.40	(−3.12, 2.39)	0.80	0.15	1.62	(−3.02, 3.33)	0.93
Immoderate *vs* sensible	3.47	1.84	(−0.14, 7.09)	0.06	2.71	1.55	(−0.32, 5.75)	0.08
Heavy *vs* immoderate	26.81	12.09	(3.11,50.50)	0.03	10.03	6.40	(−2.51,22.57)	0.12

We now compare the MI-estimates (imputation model 'without SBP') of differences in mean blood pressure with those for LD (Table 7.9). As for men, the MI-estimates for women in Table 7.9 are based on the imputed values for alcohol consumption using the first imputation model, which does not include SBP. These MI-estimates are more moderate than those for LD. The MI-estimates are all smaller than their counterparts based only on the complete records (LD). In particular, the estimated difference in blood pressure for heavy vs. immoderate drinkers is reduced to 10.03 mm Hg, and is no longer statistically significant.

We next compare these MI-estimates with their counterparts based on the imputation model including SBP (Table 7.10). As for the men, the MI-estimates 'including SBP' include all the women interviewed in 1989 ($n = 1627$). Table 7.10 shows that, compared with the results using the imputation model 'including SBP', those 'without SBP' are biased towards zero, as we

Table 7.10: Increase in SBP between successive levels of alcohol consumption for women: sensitivity to the model used for multiple imputation

	Multiple Imputation Model							
	Without SBP (n=1570)				Including SBP (n=1627)			
Contrast	**estimate**	**se**	**95% CI**	***P***	**estimate**	**se**	**95% CI**	***P***
Sensible *vs* none	0.15	1.62	(−3.02, 3.33)	0.93	−0.06	1.26	(−2.53, 2.41)	0.96
Immoderate *vs* sensible	2.71	1.55	(−0.32, 5.75)	0.08	3.12	1.52	(0.13, 6.11)	0.04
Heavy *vs* immoderate	10.03	6.40	(−2.51,22.57)	0.12	10.22	7.87	(−5.20,25.64)	0.19

found for men, but the differences between the imputation models are smaller. For example, when SBP is included in the imputation model, the MI-estimate of the mean difference in SBP for immoderate compared to sensible drinkers is 3.12 mm Hg, compared with 2.71 mm Hg when SBP is not included in the imputation model. In addition, this increase in mean SBP is estimated with greater precision, so we have a higher degree of confidence ($P = 0.04$) in the increase (95% CI 0.13 to 6.11). There is more uncertainty about the mean SBP for heavy compared with immoderate drinkers (Table 7.10: the standard error for MI 'including SBP' is 7.87, compared with 6.40 for MI 'without SBP'; $P = 0.194$ for MI 'including SBP' compared with 0.117 for MI 'without SBP').

The MI-estimates 'including SBP' indicate that we can have confidence that mean SBP for women is higher only for immoderate compared with sensible drinkers. The estimated increase in mean SBP (3.12 mm Hg, Table 7.10) is similar in magnitude to that for men (2.83 mm Hg, Table 7.8).

Rate of missing information using the two imputation models

It might be assumed that the imputation model 'including SBP' provides more information than the model 'without SBP' about the association between alcohol consumption and SBP. Table 7.11 gives the fraction of missing information (γ), as a percentage, for the MI-estimated contrasts for men and women, obtained using the 'without SBP' and 'including SBP' imputation models. As for the proportions of excessive drinkers (Table 7.2), the fraction of missing information varies between the contrasts. In Table 7.11 we see, for example, that, for men, there is more uncertainty due to missing data about the difference in SBP between immoderate and sensible drinkers ($\gamma = 42.4\%$) than about the difference in SBP between heavy and immoderate drinkers ($\gamma = 3.4\%$). Comparing the results for the two imputation models in Table 7.11, the fraction of missing information is not always smaller when SBP is included in the imputation model than when it is not. Contrasting heavy with immoderate drinkers, there is more uncertainty due to missing data when SBP is included in the imputation model ($\gamma = 32.9\%$ for

Table 7.11: Gamma (γ, %) for the contrasts between levels of alcohol consumption, for different imputation models ('without SBP' and 'including SBP')

	Men		Women	
	Imputation model		Imputation model	
	Without	Including	Without	Including
contrast	SBP	SBP	SBP	SBP
sensible *vs* none	22.0	9.6	55.5	20.4
immoderate *vs* sensible	42.4	12.9	15.0	10.5
heavy *vs* immoderate	3.4	32.9	22.2	39.4

men and 39.4% for women) than when alcohol consumption is imputed without SBP (γ = 3.4% for men and 22.2% for women). These results indicate that the inclusion of SBP in the imputation model does not give additional information about the SBP for heavy drinkers compared with immoderate drinkers.

However the observed inconsistencies could have arisen because γ is estimated. The results for γ in Table 7.11 give the fraction of information for the given set of imputations, but γ varies between imputation datasets. The results of a preliminary investigation with a simple simulated dataset (not presented) indicate that this variation is greater the smaller the number of imputations (m). Differences in γ for different models should therefore be interpreted with caution.

In this example, we have explored only the simple bivariate association between total alcohol consumption in the week and blood pressure. The 'analyst's model' for imputation would therefore use only the variables SBP and the total alcohol consumption in the diary week. The estimates of the contrasts for alcohol using this simple imputation model are predictable, since only the two variables are involved. They would be essentially the same estimates as for LD, and would increase the precision relatively little (compared to using the imputation model that includes all the other covariates) because there would be relatively little information derived from using SBP only. However, this simple imputation model is used to assess the sensitivity of our estimates of the prevalence of excessive alcohol consumption to the imputation model.

7.3.4 Sensitivity to the imputation model

Here we examine the sensitivity of MI-estimates of excessive alcohol consumption to three imputation models: 'without SBP', 'including SBP' and 'SBP only'. Table 7.12 gives the MI-estimates, standard errors and the fractions of missing information (γ) for each of these methods. (Note that the 'SBP only' model could not yield estimates of daily consumption, since it used only weekly totals: see Section 7.2.) There are no real differences between the MI-

estimates using the imputation models 'without SBP' and 'including SBP'. Nor do the standard errors differ substantially or consistently.

Table 7.12: Comparison of MI-estimates of proportions (%) consuming in excess of weekly and daily limits, and γ (as percentage): sensitivity to the model used for imputation

	Imputation model								
	'without SBP'			'including SPB'			'SBP only'		
	%	se	γ	%	se	γ	%	se	γ
Weekly									
Women									
excessive	15.9	1.20	20.4	15.9	1.12	5.9	14.6	1.21	29.0
heavy	1.0	0.37	49.3	0.8	0.28	26.2	1.6	0.50	54.7
Men									
excessive	36.6	1.54	15.1	36.8	1.61	23.8	33.8	2.15	63.6
heavy	10.4	0.96	10.2	10.2	0.95	8.8	10.6	1.28	53.6
Daily									
Women									
excessive	39.3	1.62	27.8	39.7	1.78	41.8			
heavy	13.0	1.39	53.7	12.8	1.12	25.5			
Men									
excessive	66.7	1.43	13.6	66.4	1.55	27.7			
heavy	39.6	1.49	11.6	39.9	1.53	16.8			

The MI-estimates derived using the imputation model 'SBP only', although surprisingly close to those obtained using the more complex models, tend to underestimate excessive drinking in men and women, and overestimate heavy drinking in women, relative to the estimates derived using the other imputation models. It is estimated that 33.8% of men drink excessively using the imputation model 'SBP only', compared with 36.6% for 'without SBP' or 36.8% for 'including SBP'. In contrast, for women heavy drinkers the estimate using 'SBP only' is 1.6% compared with 1.0% for 'without SBP' or 0.8% for 'including SBP'. The standard errors for the 'SBP only' model tend to be larger than the others, indicating greater uncertainty in these estimates. This is because of the greater between-imputation variance, that is, the inflation in the variance due to the missing data. The values of γ are consistently high, indicating that the fraction of missing information about the estimates is greater. This is not surprising, since the imputation used information on SBP only to impute for alcohol consumption.

7.4 Discussion

The inferences about the relationship between birthweight and blood pressure in mid-life clearly illustrate the need to deal with missing data. Taking account of missing data throws doubt on the degree of negative association of birthweight and blood pressure found using complete records only. The estimate of a 1.7 mm Hg increase in systolic blood pressure for each 1 kg decrease in birthweight agrees with the weaker association found in the meta-analysis of studies based on similar size studies (Huxley et al., 2002).

The results of the MI-estimates indicate that the native-born residents of the UK, born shortly after the end of World War II, drink rather more excessively in mid-life than would be inferred from those who completed their diaries. Over one third of men drink excessively (over the recommended limit of 21 Units of alcohol per week), and around ten percent drink heavily (over 50 Units a week). Women tend to be more moderate in their drinking than men, but even in mid-life (a time of relatively moderate drinking) around one percent of them drink heavily (over 35 Units of alcohol a week). The estimates of excess drinking over daily limits recommended in more recent health promotion literature give an even greater cause for concern. Two thirds of men drink excessively (over the recommended daily limit of 4 Units of alcohol) on at least one day of the week, while almost forty percent drink over twice this amount on at least one day. Almost forty percent of women drink excessively (over the recommended daily limit of 3 Units of alcohol) on at least one day of the week, while thirteen percent drink over twice this amount. The extent of heavy drinking has clear implications for public health. All these inferences are based on one week's drinking. If the estimates of the proportion of people drinking excessively on any one day (a maximum) were based on a longer period, they would, if anything, be greater.

As discussed in Section 5.1, comparisons with other studies are difficult because of the variation in the way the information is collected and the levels reported. Levels of drinking in the UK increased during the 1980s (Goddard, 1991) and subsequently (ONS, 1998; ONS, 2003), so that comparison needs to be with surveys contemporaneous with the 1989 NSHD interview. One survey of alcohol consumption and drinking behaviour that uses a retrospective seven day drinking diary was conducted by the OPCS Social Survey Division in 1989 (Goddard, 1991). The problem is that, even information was collected on drinking on each day of the previous week, only total consumption for the week was reported since interest at that time focussed only on drinking over the weekly limits, and not on daily excesses. Another problem with comparisons is that drinking generally declines with age and results of surveys are reported in age groups. The OPCS survey reports for the age group 35–44 years are used for comparison with our respondents at age 43. The OPCS survey reported a substantially lower prevalence of excessive and heavy total weekly drinking than we have reported above, although the same limits are used (Goddard, 1991). The prevalence of excessive drinking during the week according to the OPCS survey was only 28% for men and 9% for women; and of heavy drinking 6% for men and 1% for women. The OPCS alcohol survey clearly underestimates the extent of excessive drinking in the population compared to our results.

Increasing levels of alcohol consumption are associated with increase in systolic blood pressure (SBP). For men, this seems to hold even for those who drink at sensible levels (up to 21 Units of alcohol per week). Men who drink sensibly have on average 2.5 mm Hg higher SBP than those who drink rarely or not at all. Average increases in SBP are greater (2.8 mm Hg and 3.6 mm Hg) as alcohol consumption levels increase to immoderate (in the range 21–50 Units per week) and heavy (above 50 Units per week). For women there is no evidence that drinking at sensible levels increases blood pressure, but drinking at moderate levels is associated with a 3.1 mm Hg mean increase in SBP. The effects of heavy drinking in women cannot be estimated with reliability because of the small effective sample size, but the indications are that for some women the increase may be substantially greater than for men.

We have illustrated the value of using the multiply imputed datasets in epidemiological analyses using the example of the association of blood pressure with alcohol consumption. The analysis involves SBP, a variable that is not included in the (original) imputation model, but which is associated with alcohol consumption. The imputation model without SBP ignores the association between alcohol consumption and SBP, and as a result the estimates of the association derived from the analysis using this model are biased towards zero. However the extent of the bias is small because the other variables involved in the imputation model account for much of the association between SBP and alcohol consumption. Adding SBP to the imputation model adds little extra information about alcohol consumption, and the estimates of alcohol consumption are not affected by using this more comprehensive imputation model. The problem posed by the use of the imputations in future analyses is mitigated, in the same way as the possibility of MNAR, by virtue of the richness of the set of covariates available for the imputation of alcohol consumption. Using only SBP in the imputation increases the uncertainty in the estimates of excessive alcohol consumption and, by restricting the covariates used, exposes the data to a greater threat of MNAR.

Imputing using only the variables in the model for analysis is a simple approach which evades investigation of the missing data problem itself. This approach has been shown to be adequate in the estimation of the dependence of blood pressure on birthweight, but this is the case only because the missing data is in a covariate that is not a confounder in the analysis. In general, it has serious deficiencies. It does not exploit the rich set of covariates so that the possibility of MNAR (which is relative to that set of covariates) is greater. Secondly, it is does not allow the exploitation of the results for other purposes. For example, the imputation of alcohol consumption using only SBP yields very uncertain estimates of weekly alcohol consumption, and cannot be used to explore other aspects of alcohol consumption, such as excessive drinking on a daily basis, which may be relevant to epidemiological studies.

Chapter 8

Discussion

At the turn of the twenty first century, excessive alcohol consumption is widely acknowledged as a public health problem. Alcohol misuse now costs the NHS up to Â£1.7 billion a year, UK plc loses Â£6.4bn in lost productivity, while the cost in related crime and public disorder is estimated to be up to Â£7.3bn (Strategy Unit, 2003). Epidemiological studies of the adverse consequences of excessive alcohol consumption require information on the alcohol consumption of individuals in the general population. However, whether an individual drinks excessively is difficult to ascertain, because alcohol consumption varies within individuals over time. Chronic and serious adverse health consequences such as cirrhosis or CHD result from long-term excess consumption and are more closely linked to excess in total consumption over a period (Rehm et al., 1996). Social consequences (such as domestic violence, accidents, or sickness absence from work) are thought to be more closely related to excessive consumption on a single occasion (Rehm et al., 1996). Hence not only total quantity, but also patterns of drinking, or the distribution of excessive drinking over occasions, have recently been recognised as being important in alcohol research (Grant and Litvak, 1997). Distinguishing between these aspects of alcohol consumption requires the collection of detailed data, such as found in a daily diary. Such data has rarely been collected in large-scale studies of the general population, but in 1989 this was collected by the MRC National Survey of Health and Development (NSHD) in a diet diary. This data provides the opportunity for epidemiological studies of alcohol consumption.

As discussed in Chapter 3, as well as the seven-day diary, the survey also collected data simply on the total numbers of different alcoholic beverages taken in the past week (weekly recall). Respondents to both instruments (seven-day diary and weekly recall) recorded substantially lower consumption in the weekly recall than in the seven-day diary, in which they recorded all food and drink taken on each day over a week. Although weekly recall is a much simpler way of collecting information about total alcohol consumption in a week, if this instrument is used excessive drinking would be substantially underestimated. Respondents also answered the CAGE questionnaire about their problems with drinking, both ever and in the past year. The extent of underestimation in the weekly recall summary measure was greater for those who had drink problems in the past year. This implies that the weekly recall alone could not give good information about the total amount of alcohol consumed by respondents during a week. Besides it cannot collect as detailed information as a diary can. It cannot tell us about the way in which respondents' drinking was distributed over the week, about their pattern of drinking. Only the diary data could do this. On the one hand the detailed information the diary gives is essential for alcohol research; on the other, its recording requires a serious commitment from respondents

with the consequence that many of the respondents did not complete all seven days of their diary. The diary has a substantial proportion of missing data.

8.1 The importance of dealing with missing data on alcohol consumption

Missing data has always been a problem for epidemiology, particularly in longitudinal studies. In the past, missing data in epidemiological studies was dealt with mainly by naïve approaches (such as ignoring cases with missing data, or using 'mean value replacement'). In general, epidemiologists rely on software packages to analyse data and the methods that are used in practice are restricted to those which are provided in the available software. Attention to missing data has increased in recent years. Multiple imputation (MI) has been shown to have good statistical properties (Little and Rubin, 1987), but it has been little used in practice, because the software to implement it has not been available until recently (Zhou et al., 2001). This is especially true in alcohol research (Figueredo et al., 2000; Gmel et al., 2001).

Most standard software packages automatically exclude cases with incomplete records from the analysis, without warning the user (Figueredo et al., 2000). As more variables are included in an analysis, the potential for excluding cases increases. The example of the regression coefficient of systolic blood pressure on birthweight, discussed in Chapter 4, illustrated this problem. It was shown that excluding cases with incomplete records leads to a confounding of the effect of selecting only complete cases with any modifying effect of a covariate in the regression model. The selection effect may be negligible when only a small proportion of subjects have missing data on the covariate, as when blood pressure was considered in relation to birthweight, allowing for gender and social class, but without taking alcohol consumption into account. However, the coefficients in a regression analysis can be particularly sensitive to a few outliers which have a strong influence. When total alcohol consumption over the diary week was added as a covariate in the regression model, leading to exclusion of a large proportion of cases with incomplete records, the effect was substantial. Comparing the results of analyses with and without the covariate, but using the same set of cases (those with complete records), showed relatively small differences, indicating that the substantial change in the coefficient was due to the effect of selecting only respondents who had complete dairies.

Exclusion of cases with incomplete records from the analysis yields unbiased estimates only when the data is missing completely at random (MCAR). This is rarely the case in surveys, and we have seen (in Chapter 5) that alcohol consumption in the NSHD is no exception. Those from lower social class, of origin and in adulthood, and with lower educational qualifications and lacking in basic literacy or numeracy skills, were less likely to complete their diaries. Those with higher drinking according to their weekly recall, those with higher CAGE scores, and smokers were also less likely to complete their diaries, and at the same time to have higher alcohol consumption in their diary. However, this need not spell disaster for estimating alcohol consumption if we can take into account observed information which is related to alcohol consumption and to missingness, to predict the missing values on alcohol consumption.

Imputation of missing data values enables us to exploit the observed information related to alcohol consumption by assuming that it is related to the unobserved values in the same way as it related to the observed values, that is, that the data is missing at random (MAR) conditional on the observed data.

Important observed information relevant to alcohol consumption is contained in the weekly recall, CAGE score, gender, and social class, variables for which the proportion of missing data is relatively small. Additional information about drinking on any particular day can be derived from recorded days, most respondents having completed at least the first two days of the diary. We also know from those who completed their diaries that the pattern of drinking varied over the days of the week, and the pattern depends on gender and social class. Of importance to public health is the finding that the more days people drink, the more they tend to drink on each day. Higher social classes tend to drink more regularly (on more days of the week) than lower social classes. Men in higher social classes drank more moderately on each day than those in lower social classes; it seems they were more responsive to health messages. But for women this was not the case: higher social class women drink more often, and to drink larger amounts, than those from lower social classes, as found in other European countries (Makela, 1999).

8.2 Methods for dealing with missing data on alcohol consumption

The object of any analysis is to make valid inferences. The inferences should be unbiased and efficient. One component of the uncertainty is due to missing values, and this has to be reflected in all inferences. Deterministic or single imputation methods cannot do this since an analysis based on the (singular) completed dataset treats missing values as if they were observed. A method that can make full use of the observed data in the partially complete records will be more efficient. Some methods can only do so to a limited extent because they do not use all this information at one time. The preferred methods are those based on maximum likelihood, such as the EM algorithm, since these provided the most efficient estimators of the parameters in the statistical models for the relationships between variables. In addition to uncertainty arising from missing values, valid methods should take into account the uncertainty in the estimation of parameter values. Of those examined, the only procedures that meet all these criteria were Schafer's. Some methods do not preserve the association between variables (SOLAS Propensity Score), and are therefore unsuitable for epidemiological applications.

A pervading problem when dealing with missing data (other than that missing by design) is that we do not know why the data is missing or what it might have been, and therefore do not know whether our inferences are unbiased. Hence the bias in a method was evaluated by using simulated data (Chapter 6). The simulated data consisted of the cases in the NSHD database that had complete records (diary completers), in which some of the records were artificially set as "missing" according to known mechanisms of missingness. The semicontinuous distribution of alcohol consumption (either a definite zero, or a log-Normally distributed positive quantity) poses a technical problem since statistical procedures for continuous data are generally not

robust to such extreme departures from Normality. To overcome this problem, sign and positive amount were dealt with separately. This approach could only be implemented using software that includes procedures for imputing both categorical and continuous missing data. The only methods that were unbiased, even when the data was MCAR, were those using Schafer's procedures for MI. The diary data was re-ordered by day of the week to preserve the patterns in alcohol consumption over the week. The chosen method was based on Schafer's procedures for MI: CAT (for sign) and NORM (for positive quantity). Schafer's procedure MIX (designed for a mixture of categorical and continuous variables) could be used to impute sign and quantity simultaneously. However, using MIX was not practical in this application since the complexity of the model made it impractical to allow for the uncertainty in the parameters using Data Augmentation with the MIX procedure.

The question of validity of the inferences also depends on the MAR assumption holding. MI has been said to provide protection against the data being missing not at random (MNAR) provided a rich enough set of covariates is used. This was demonstrated to hold with the chosen method since the bias in the estimates was small even when the data was MNAR.

The multiple (five) completed datasets can be used in future analyses of the NSHD that involve alcohol consumption, without the analyst being distracted by the missing data in this variable. Having found the best method, from theory and which worked in practice, it was applied to the NSHD data, to give estimates of levels of excessive alcohol consumption in middle life in the population born in the immediate post-war years. The levels of excessive and heavy alcohol consumption are based on the government-recommended sensible drinking guidelines for weekly and for daily alcohol consumption. The results indicate that a substantially higher proportion of this population drink excessively or heavily, based on recommended daily limits, than appear to do so when their drinking is averaged over the whole week, particularly for women drinking over twice the recommended daily limit (over 6 Units), described as bingeing (Strategy Unit, 2003). The level of binge drinking in Britain, particularly amongst young women, has been a recent cause for concern. Our estimates show that recent reports of excessive drinking are underestimated. It is reported that almost one in four women now drink more than the recommended daily alcohol limit of three units at least once a week, and one in ten women now 'binge drink' (that is to say they consume six units of alcohol in one session—equivalent to two thirds of a bottle of wine) at least once a week (Asthana and Doward, 2003). The corresponding estimates from the NSHD of 39% and 13% are not only greater, but relate to women in mid-life, when drinking, and particularly binge drinking declines, and to 1989, since when it is known that levels of drinking in the whole population have increased (Strategy Unit, 2003).

A problem raised by Fay and others (Fay, 1992; Meng, 1994) is that using the imputations in this way in future analyses would not be valid if the analyst's model includes a variable that has not been included in the imputation model. The example used to illustrate this was the analysis

of birthweight against systolic blood pressure (SBP). The effect of this was assessed by testing the sensitivity of inference to the original imputation model, and to the same one including SBP in addition. The estimates were, as predicted, were biased towards zero. However this effect was small because SBP did not add (much) information about alcohol consumption, in addition to the other covariates. The estimates of excessive alcohol consumption were essentially unchanged when SBP was included in the imputation model compared with when it was not. The implication of this is that if the set of covariates used in the imputation model is rich enough then the imputations will be robust enough to include relationships with additional variables used in future analyses.

8.3 Implications of this thesis for epidemiological methods

The bulk of the mortality and morbidity in the developed world is attributable to chronic diseases, which generally have a long period of development. Thus longitudinal studies of the life course are becoming increasingly important in epidemiology (Ben-Shlomo and Kuh, 2002; Kuh and Ben-Shlomo, 1997) because they provide information about progressive pathways to disease (Hardy and Wadsworth, 2001). There is also a growing interest in the early life origins of disease in later life, for example, in the foetal origins of coronary heart disease (Barker, 1995), which can be studied only with long term longitudinal studies like the MRC National Survey of Health and Development (Hardy et al., 2003). Missing data is a particular problem in longitudinal studies because of the effect of attrition. Some people in the original sample drop out of the study altogether, and the number of these dropouts generally increases as the study progresses. However, longitudinal studies have the potential to provide information about people who have dropped out of the study from their responses at earlier sweeps. The longer the study, the greater the problem of attrition, but the more we know about those who have dropped out. In a cross-sectional survey there will be no information about case non-response except for the prior known criteria of sample selection, usually demographic variables. A strength of longitudinal surveys is that we have information about people measured before they drop out.

The methods investigated in this dissertation for item non-response in the diet diary are generally applicable to case non-response due to attrition in longitudinal studies. The problem of attrition in longitudinal studies is logically no different from the problem of item non-response. From a longitudinal perspective, case non-response may be viewed as a set of item non-responses in a longitudinal record. Case non-response is often dealt with using weighting, while imputation is used for item non-response. But case non-response in longitudinal studies can be dealt with just as item non-response, using imputation methods (Rovine and Delaney, 1990). The only difference is that the information about dropouts must be derived from variables measured at some earlier point in time. If the strength of association between variables diminishes with separation in time, variables measured at earlier sweeps may be less informative about case non-response than are concurrent variables about item non-response.

8.3.1 Implications for collection of data on alcohol consumption

Background data is crucial to provide imputations that enable valid inferences to be made. It mitigates the consequences of the data being MNAR and it protects against bias in future analyses that may use other variables. This has implications for the ways in which data on alcohol consumption is collected. It underlies the importance of collecting the data in different ways, as was the case with the NSHD.

The recalled weekly total alcohol consumption (Appendix 1, question 1) may not be as valid as the diary itself as a direct source of information on alcohol consumption, but it provides valuable background information about people's level of drinking, and is more likely to be completed. Also, it is important that at least two days of the diary data was available for the vast majority of respondents (96%). Since multiple imputation uses the information in partially completed records of the diary, the information obtained in the first two diary days was invaluable. The availability of this data was ensured by using a suitable method of data collection. In the MRC NSHD 1989 survey the nurse interviewer collected and copied the information on the first two days of the diary before leaving the diary with the respondents to complete themselves. The ideal way of collecting the diary data to ensure that the data were complete would have been for the interviewer to call on the respondent on each day of the week. The survey strategy represents an effective economic compromise in what was potentially a very expensive way of collecting such detailed information as is obtained from a daily diet diary.

In dealing with the data on alcohol consumption we have seen that zero may often be confused with missing (Section 3.2.5). Respondents tend not to answer questions that they feel have no relevance to them, such as those asking how many drinks they had had in the previous week when in fact the respondent is an abstainer or does not drink the particular type of beverage. Reporting on the 1989 OPCS Social Survey Division study of alcohol consumption, Goddard (1991) states *'a considerable number of informants gave as their reason for not wishing to take part in the survey the fact that they did not drink at all and could not be persuaded of its relevance to them.'* The evidence that individuals tend to complete only options within questions that apply to them and their positive behaviour is a general problem for collecting survey data and should be born in mind when interpreting the data collected (Dengler et al., 1997). Designing questionnaires to avoid such questions can help reduce the extent of missing data.

Alcohol surveys generally first ask the respondent if they drink at all or ever take alcohol, and then direct those who answer in the affirmative to questions about drinking, such as about quantities of alcoholic beverages. Designing the recall question ('how many of the following drinks have you had?', Appendix 1, Q1) in this way would reduce the non-response to this question. Such a change in the question would have the potential to identify abstainers. The identification of abstainers, as opposed to those who simply did not drink during the period, would provide useful information for epidemiological research. (It would also have

implications for the method of imputing the diary. Known abstainers would have to be removed from the data before imputation, as it would be known for certain that they did not drink.) The diary, however, is not an instrument specifically for the collection of alcohol data. The collection of data on alcohol consumption is simply part of the reporting of all food and drink consumed. The advantage of this seems to be that it alleviates the potential problem of non-response by heavy drinkers caused by embarrassment about their drinking. Collecting information on alcohol consumption using the diet diary data implies that the alcohol data is less likely to be MNAR.

8.3.2 Implications on the use of procedures for imputation

Although multiple imputation (MI) may be theoretically ideal, the method for dealing with missing data in practice depends on the nature of the data, and the extent of missing data. In this application to item non-response to alcohol consumption, the important features are that there is a large amount of missing data but also a rich set of covariates to provide good auxiliary information. This makes the effort required for MI worthwhile. The use of the good auxiliary information increases efficiency (reduces the uncertainty about the missing values), while the use of 'proper' imputation methods ensures that the differences between the sets of plausible values reflect the uncertainty about the missing values. The methods have been shown to be unbiased when the data is missing with a probability increasing with the weekly recalled amount (MAR conditional on weekly recall). They also provide some protection against MNAR. Finally, the imputed datasets can be used in any analysis.

When the proportion of missing data is relatively small, simpler methods may be adequate. For example, in the imputation model one covariate was 'adult social class' (Section 5.2). For this variable current social class was used and the missing values filled in on a last value carried forward basis, as is common practice. If the respondent was unemployed at the time of the 1989 survey, the social class of the most recently available occupation, when the respondent was aged 36 or 26 years, was used. Only 1.8% (60: 13 men and 47 women) of respondents did not have a social class defined in this way. Of these, 42 women were assigned the social class of their spouse, and the remaining 18 respondents were assigned a social class based on parental occupation, likewise based on the most recently available information. There may be some uncertainty about whether the most recently available social class is applicable, and we should use MI for the missing values, but where such a small proportion of data is missing, it would not affect inferences. Examination of interview sheets themselves can also provide valuable information about the nature of the missing data, for example, in those missing one of the items in the recall (Section 3.2.5). This partial non-response applied to a larger proportion (22%) of respondents, but the evidence from the interview sheets and the data provided such compelling evidence that the blanks should have been zeros, that they were deterministically imputed as such. When they were, more correctly, multiply imputed (Longford et al., 2000), the impact on inferences was only slight.

The choice of procedure may be limited by available software and, depending on the distribution of the variable to be imputed and the covariates used, using a simpler, but not ideal, method is better than ignoring cases with incomplete data. MI of a Normally distributed continuous variable could be implemented using repeated runs of SPSS Regression. This is not ideal because the procedure uses a series of separate regressions for each variable to be imputed, and so is less efficient than procedures based on maximum likelihood. SPSS EM cannot be recommended however, even for Normally distributed variables. Not only does it give biased estimates, but also it is a deterministic procedure, and so it can only be used for single imputation and fails to account for the uncertainty about missing values. Using this procedure would give a false confidence in estimates that are biased, which make it more likely that false conclusions will be drawn. The SOLAS model based method for continuous data uses separate regressions on the data sorted as monotone missing, so has the same drawbacks as SPSS Regression (less than maximum efficiency), but has the flexibility that different covariates can be used for each variable to be imputed. Like SPSS Regression it would be better than naïve methods if the variable to be imputed is approximately Normally distributed. The SOLAS Discriminant Method can only be recommended for imputing variables which are nominal, provided the covariates are continuous. If the covariates are categorical and skewed, this procedure gives biased results and, even with categorical covariates that are not skewed, it distorts the relationships between the covariate and the variables to be imputed.

The strategy of separating the sign and the positive quantity to avoid the problem presented by the semicontinuous distribution of alcohol consumption could be applied to any semicontinuous variable. Variables with a semicontinuous distribution are quite commonly encountered, for example: number of cigarettes smoked, numbers of times a Class A drug is used, numbers of days spent abroad in a year, expenditure on consumer products (for example a TV, or vitamin supplements). If there is only one semicontinuous variable to impute, Schafer's MIX would be appropriate, as in the imputation of total alcohol consumption for the diary week (Section 7.2).

As MI becomes more readily available in standard software, it may become routine practice to impute at the point of analysis. This may lead analysts to impute using only the variables in their model for analysis. That is, the dataset used for the analysis is 'augmented' based on MAR conditional on the variables in that model. This approach ensures that the relationships amongst the variables in the model for analysis are included in the model for imputation. The disadvantages of this approach are that a limited set of covariates may not protect against MNAR, the inferences could be less efficient (because they may have omitted important information about a variable), and the imputations will not be useful in other analyses. This approach could be used when the extent of missing data is more limited, or the background data unavailable. The example using only SBP to impute alcohol consumption illustrates that it is better to use a simple imputation model and MI than to ignore missing data.

Multiple imputation may offer a good solution to the problem of missing data, but the validity of inferences from an analysis using imputed data depends on the validity of the assumptions, that is, on the imputation model. Sensitivity analysis is needed to examine the effect of different assumptions on the inferences. In the words of Horton and Lipsitz (2001): *'The existence of software that facilitates its [multiple imputation] use requires the analyst to be careful about the verification of assumptions, the robustness of imputation models, and the appropriateness of inferences.'*

8.4 Limitations of this dissertation

Software for multiple imputation is currently being developed. This thesis has not been able to consider all the software available now (at the end of 2003). The current release of the software package SAS (SAS/STAT Software, 2003), widely used in the pharmaceutical industry, includes MI procedures based on those of Schafer used in this work. SAS PROC MI could be used to implement the method developed here. MLwiN does not have a procedure specifically for imputation. However, it includes a facility for performing MCMC sampling on a model for data (Browne, 2003), and a skilled programmer could use this to do MI on multivariate Normal or mixed categorical and multivariate Normal data. Similarly, WinBUGS is specially designed to carry out Bayesian inference on complex models using MCMC methods. In WinBUGS, the missing values are treated like parameters, and the software samples from their Bayesian posterior distribution conditional on the observed data. However extensive programming skills are required to implement this. This puts approaches based on MLwiN or WinBUGS out of reach for most epidemiologists. STATA, widely used in epidemiology, does not have MI procedures as part of its main release. Most reported use of STATA for imputation seems to involve Hot Deck or Mean Score methods (Mean Score is a deterministic version of Hot Deck so could not be used for MI). However Carlin (2003) has developed STATA procedures for combining MI datasets produced using other software (and they have been used for example by De Stavola et al. (in press)). Other software resources with varying capability for imputation, reviewed by Horton and Lipsitz (2001), include MICE (a library for S-Plus and R), Amelia, IVEware, HLM and LISREL.

The method adopted for dealing with semicontinuous data in two steps is not ideal. It fails to relate a particular sign with its associated positive quantity: that is, whether someone drinks on a particular day cannot be associated with how much they drink (a particular sign to its positive amount). Algorithms designed to cope with semicontinuous data in this way are currently under development (Olsen and Schafer, 1998) but are not yet generally available. The problem of dealing with the modelling and imputation of semicontinuous variables is the subject of ongoing research (Schafer and Olsen, 1999).

This thesis has not dealt with attrition from the birth cohort. The reference population is considered to be that well represented by the sample consisting of those interviewed in 1989. The NSHD uses simple, but effective, ways to maintain contact and improve case response.

Study members are sent a birthday card every year, including feedback on the findings of the study, which informs the members of their important contribution to research (Wadsworth et al., 2003; Wadsworth et al., 1992). Heavy drinkers would be underrepresented in this sample, for example, because of the association of alcohol abuse with homelessness (Reardon et al., 2003). Unfortunately, variables available from early sweeps, when contact was maintained with more of the original birth cohort, were not strongly associated with alcohol consumption in mid-life. So no background information is available as a basis for imputation. Relevant information, such as parental drinking, was not collected, as interest in alcohol consumption only developed in the 1980s. One disadvantage of this cohort study (or any long term longitudinal studies) is that the information collected depends on the secular scientific interest (Wadsworth et al., 2003). However, the social class on which the sample was stratified could be used as a basis for the imputation of a summary measure of alcohol consumption, and future work could include a sensitivity analysis to the inclusion of all those in the original birth cohort.

8.5 Moving forward

Dairy data supplemented by properly generated plausible values enable us to move forward in alcohol research. 'To move forward, we need to be able to describe drinking patterns more accurately.' (Grant and Litvak, 1997). The reviewers of Grant and Litvak's book repeatedly refer to the need for more and better data to be able to describe drinking patterns (Grant and Single, 1997). Detailed daily data is needed rather than just weekly totals, and this has rarely been collected in large samples of the general population. 'The social dimension has not been properly investigated because of a lack of large-scale cohort studies with adequate measures of social variable.' (Rehm et al., 1996).

Using the analytical developments in this thesis, the information in alcohol diary data can be exploited fully and efficient inferences made about different aspects of alcohol consumption. This would facilitate future research on how pattern of drinking (such as frequency of drinking and frequency of bingeing), rather than average weekly drinking, is associated with, for example, hospital admissions for injury, frequency of visits to the doctor, sickness absence from work (weekday bingeing), or frequency of change of employment.

The study of the application of multiple imputation in this thesis highlights important issues for epidemiology in the future. Missing data is ubiquitous in epidemiological studies. Until recently the standard approach to missing data in epidemiology has been to use naïve methods that are likely to result in inefficient or invalid inferences. Multiple imputation is a universal method of dealing with missing data which can yield efficient inferences which are valid, provided the assumptions are met. Schafer's procedures for multiple imputation provide the epidemiologist with the software to deal with the problem of missing data properly.

However, this study has highlighted some issues which would be important in any application, and which the epidemiologist should not ignore. Any method of dealing with missing data is

dependent on assumptions which may not hold, and some sensitivity analysis of the inference to these assumptions is called for. The naïve methods which have typically been used until recently make these assumptions implicitly and the analyst may not be aware of this. Yet because multiple imputation has focussed attention on missing data problems and has made the assumptions explicit, the approach has been the subject of criticisms that apply equally to any method of dealing with missing data. Addressing these criticisms serves to increase our awareness of the problems posed by missing data.

This thesis has demonstrated that naïve approaches, and the inappropriate use of software for imputation, can result in inefficient and biased inferences that would lead the epidemiologist to erroneous conclusions, and that Schafer's procedures for multiple imputation offer a means for epidemiology to move forward.

The work undertaken in this thesis can be appropriately summarised in the words of Barnard and Meng (1999), discussing multiple imputation:—

> *'Cautions are needed, however, just as with any statistical methodology. It is clear that if the imputation model is seriously flawed in terms of capturing the missing-data mechanism, then so will be any analysis based on such imputations. This problem can be avoided by carefully investigating each specific application, by making the best use of knowledge and data about the missing-data mechanism, and by performing various model checking procedures, in particular, posterior predictive checks. This is not an additional burden for using Rubin's method, but rather a fundamental requirement for any general method that attempts to produce statistically and scientifically meaningful results in the presence of incomplete data.'*

References

Alanko, T. (1981) An overview of techniques and problems in the measurement of alcohol consumption. In *Research Advances in Alcohol and Drug Problems* (eds R. G. Smart et al.), **8**, 209–226. New York: Plenum Press.

Altman, D. G. (1991) *Practical Statistics for Medical Research.* London: Chapman & Hall.

Aquilino, W. S. (1992) Telephone versus face-to-face interviewing for household drug use surveys. *International Journal of the Addictions*, **27**, 71–91.

Arnold, A. M. and Kronmal, R. A. (2003) Multiple imputation of baseline data in the cardiovascular health study. *Americam Journal of Epidemiology*, **157**, 74–84.

Arria, A. M. and Gossop, M. (1998) Health issues and drinking patterns. In *Drinking Patterns and Their Consequences* (eds M. Grant and J. Litvak), International Center for Alcohol Policies, 63–87. Washington, DC: Taylor & Francis.

Asthana, A. and Doward, J. (2003) Binge drinking: do they mean us? Special Report, Medicine and Health, *The Observer*, 21st December, 2003

Bamford, J., Sandercock, P., Dennis, M., Burn, J. and Warlow, C. (1990) A prospective study of acute cerebrovascular disease in the community: the Oxfordshire Community Stroke Project — 1981-86. 2. Incidence, case fatality rates and overall outcome at one year of cerebral infarction, primary intracerebral and subarachnoid haemorrhage. *Journal of Neurological and Neurosurgical Psychiatry*, **53**, 16–22.

Barker, D. J. P., ed. (1992) *Fetal and Infant Origins of Adult Disease.* London: BMJ Publishing Group.

Barker, D. J. P. (1994) *Mothers, Babies, and Disease in Later Life.* London: BMJ Publishing Group.

Barker, D. J. P. (1995) Fetal origins of coronary heart disease. *British Medical Journal*, **311**, 171–4.

Barnard, J., and Meng, X.-L. (1999) Applications of multiple imputation in medical studies: from AIDS to NHANES. *Statistical Methods in Medeical Research*, **8**, 17–36.

Ben-Shlomo, Y. and Davey Smith, G. (1991) Deprivation in infancy or in adult life: which is more important for mortality risk? *The Lancet*, **337**, 530–4.

Ben-Shlomo, Y. and Kuh, D. (2002) A life course approach to chronic disease epidemiology: conceptual models, empirical challenges, and interdisciplinary perspectives. *International Journal of Epidemiology*, **31**, 285–293

Bongers, I. M. B., van de Goor, L. A. M., van Oers J. A. M. and Garretsen, H. F. L. (1998) Gender differences in alcohol-related problems: controlling for drinking behaviour. *Addiction*, **93**, 411–421.

Braddon, F. E. M, Rodgers, B., Wadsworth, M. E. J. and Davies, J. M. C. (1986) Onset of obesity in a 36 year birth cohort study. *British Medical Journal*, **293**, 299–303.

Braddon, F. E. M, Wadsworth, M. E. J, Davies, J. M. C. and Cripps H. A. (1988) Social and regional differences in food and alcohol consumption and their measurement in a national birth cohort. *Journal of Epidemiology and Community Health*, **42**, 341–349.

Brand, J., van Buuren, S., van Mulligen, E.M., Timmers, T. and Gelsema, E. (1994) Multiple imputation as a missing data machine. In *Proceedings of the Eighteenth Annual Symposium on Computer Application in Medical Care (SCAMC)* (ed. J. G. Ozbolt), 303–6. Philadelphia: Hanley & Belfus, Inc.

Breeze, E. (1985) *Women and Drinking*. OPCS Social Survey Division. London: HMSO.

Browne, William J. (2003) *MCMC Estimation in MLwiN*. University of London Institute of Education, Centre for Multilevel Modelling. Version 2.0 July 2003.
Available for free download from:
http://www.maths.nott.ac.uk/personal/pmzwjb/materials/mcmcman2.pdf

Caetano, R. and Clark, C.L. (1998) Trends in alcohol cnsumption patterns among whites, blacks and Hispanics: 1984–1995. *Journal of Studies on Alcohol*, **59**, 659–668.

Carlin, J. B., Li, N., Greenwood, P. and Coffey, C. (2003) Tools for analyzing multiple imputed datasets. *The STATA Journal*, **3**, 215–216.
Abstract: http://www.stata-journal.com/abstracts/st0042.pdf
Full text for subscribers: http://www.stata-journal.com/software/sj3-3/

Christensen, K., Vaupel, J. W., Holm, N. V. and Yashin, A. I. (1995) Mortality among twins after age 6: fetal origins hypothesis versus twin method. *British Medical Journal*, **310**, 432–6.

Cox, B. D., Huppert, F. A. and Whichelow, M. J. (1993) In *The Health and Lifestyle Survey: Seven years on*. Dartmouth Publishing Co Ltd, Aldershot, UK.

Crawford, A. (1986) A comparison of participants and nonparticipants from a British general population survey of drinking practices. *Journal of the Market Research Society*, **28**, 291–297.

De Lint, J. (1981) 'Words and deeds': responses to Popham & Schmidt. *Journal of Studies on Alcohol*, **42**, 359–360.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Ser. B, **39**, 1–38.

Dengler, R. (1996) Smoking and alcohol consumption in Trent, UK: an analysis of item non-response. *Journal of Epidemiology and Community Health*, **50**, 687.

Dengler, R., Roberts, H. and Rushton L. (1997) Lifestyle surveys — the complete answer? *Journal of Epidemiology and Community Health*, **51**, 46–51.

De Stavola, B.L., dos Santos Silva, I., McCormack, V., Hardy, R.J., Kuh, D.J, and Wadsworth, M. E. J. (in press) Childhood growth and breast cancer. *American Journal of Epidemiology*.

Dight, S. E. (1976) *Scottish Drinking Habits*. OPCS Social Survey Division. London: HMSO.

Dillner, L. (1955) News: Colleges call for safe drink limits to stay. News, *British Medical Journal*, **310**, 1623.

Dos Santos Silva, I., de Stavola, B. L. , Mann, V., Kuh, D., Hardy, R., and Wadsworth M. E. J.(2002) Prenatal factors, childhood growth trajectories and age at menarche. *International Journal of Epidemiology*, **31**, 405–12.

Duffy, J. C. (1993) Alcohol consumption and control policy. *Journal of the Royal Statistical Society* Ser. A, **156**, 225–230.

Duffy, J.C. and Alanko, T. (1992) Self-reported consumption measures in sample surveys: a simulatiuon study of alcohol consumption. *Journal of Official Statistics*, **8**, 327–50.

Edwards, G. (1994) *Alcohol Policy and the Public Good.* Oxford University Press.

Edwards, G., Chandler, J. and Hensman, C. (1972) Drinking in a London Suburb. *Quarterly Journal of Studies on Alcohol.* Supplement, **6**, 69–93.

Ely, M., Hardy, R., Longford, N. T. and Wadsworth, M. E. J. (1999) Gender differences in the relationship between alcohol consumption and drink problems. *Alcohol and Alcoholism*, **34**, 894–902.

Ewing, J. A. (1984) Detecting alcoholism, the CAGE questionnaire. *Journal of the American Medical Association*, **252**, 1905–07.

Faculty of Public Health Medicine (1996) Preventing the harm related to alcohol use; reducing population risk. *Guidelines for Health Promotion, No. 46.* London: Policy Office, Faculty of Public Health Medicine.

Faris, P. D., Ghali, W. A., Brant, R., Norris, C. M., Galbraith, P. D. and Knudtson, M. L. (2002) Multiple imputation versus data enhancement for dealing with missing data in observational health care outcome analyses. *Quarterly Journal of Studies on Alcohol*, **55**, 184–191.

Fay, R. E. (1992) When Are Inferences from Multiple Imputation Valid? In *Proceedings of the section on Survey Research Methods*, American Statistical Association, 227–232.

Fay, R. E., (1996) Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association*, **91**, 490–498.

Figueredo, A. J, McKnight, P. E., McKnight, K. M. and Sidani, S. (2000) Multivariate modelling of missing data within and across assessment waves. *Addiction*, Supplement 3: *State of the Art Methodologies in Alcohol Related Health Sciences Research*, S361–S380.

Gilks, W. R., Richardson, S and Spiegelhalter, D. J. (1996) *Markov Chain Monte Carlo in Practice.* London: Chapman & Hall.

Gmel, G. (2001) Imputation of missing values in the case of a multiple item instrument measuring alcohol consumption. *Statistics in Medicine*, **20**, 2369–2381.

Goddard, E. (1991) *Drinking in England and Wales in the late 1980s.* OPCS Social Survey Division. HMSO, London.

Goyder, J. (1987) *The Silent Minority—Non Respondents on Sample Surveys.* Cambridge: Polity Press.

Graham, K., Wilsnack, R., Dawson, D. and Vogeltanz, N. (1998) Should alcohol consumption measures be adjusted for gender differences? *Addiction*, **93**, 1137–1147.

Grant, M. and Litvak, J. (1997) Introduction: beyond per capita consumption. In *Drinking Patterns and Their Consequences* (eds M. Grant and J. Litvak), International Center for Alcohol Policies. Taylor & Francis, USA.

Grant, M. and Single, E. (1997) Shifting the paradigm: reducing harm and promoting beneficial patterns. In *Drinking Patterns and Their Consequences* (eds M. Grant and J. Litvak), International Center for Alcohol Policies. Taylor & Francis, USA.

Greenfield, T. K., Midanik, L. T. and Rogers J. D. (2000) A 10-year national trend study of alcohol consumption, 1984-1995: is the period of declining drinking over? *American Journal of Public Health*, **90**, 47–52.

Greenland, S. and Finkle, W.D. (1995) A critical look at methods for handling missing covariates in epidemiologic regression analyses. *American Journal of Epidemiology*, **142**, 1255–64.

Groves, R. M. (1989) *Survey Errors and Survey Costs*. New York: Wiley.

Hajema, K.-J. and Knibbe, R. A. (1998) Changes in social roles as predictors of change in drinking behaviour *Addiction*, **93**, 1717–1727.

Harding, S., Brown, J., Rosato, M. and Hattersley, L. (1999) Socio-economic differentials in health: illustrations from the Office for National Statistics Longitudinal Study. *Health Statistics Quarterly*, Spring, 1999. Office for National Statistics.

Hardy, R., Kuh, D., Langenberg, C. and Wadsworth, M. E. J. (2003) Birthweight, childhood social class, and change in adult blood pressure in the 1946 British birth cohort. *The Lancet*, **362**, 1178–1183.

Hardy, R. and Wadsworth, M. E. J. (2001) The British Birth Cohort Studies: Childhood influences on adult life. *American Statistical Association's 2000 Proceedings of the Section on Government Statistics and Section of Social Statistics*, 28–34.

Hawkins, J.D., Graham, J.W., Maguin, E., Abbott, R., Hill, K. G. and Catalano, R. F. (1997) Exploring the effects of age of alcohol use initiation and psychosocial risk factors on subsequent alcohol misuse. *Journal of Studies on Alcohol*, **58**, 280–90.

Health Education Council (1984) *Annual Report 1983-4*. Health Education Council, London

Health Education Council (1985) *'That's the Limit.'* Health Education Council, London

Health Education Authority (1996) *Think about drink*. Health Education Authority, London

Heath, A. C., Howells, W., Kirk, K. M., Madden, P. A., Bucholz, K. K., Nelson, E. C., Slutske, W. S., Statham, D. J. and Martin, N. G. (2001) Predictors of non-response to a questionnaire survey of a volunteer twin panel: findings from the Australian 1989 twin cohort. *Twin Research*, **4**, 73–80.

Hedges, B. (1996) Alcohol Consumption. In *Health Survey for England, 1994*, (eds Colhoun, H. and Prescott-Clarke, P.) 338–365. London: HMSO.

Heitjan, D. F. and Little, R. J. A. (1991) Multiple imputation for the fatal accident reporting system. *Applied Statistics*, **40**, 13–29.

Hope, S., Power, C. and Rodgers, B. (1998) The relationship between parental separation in childhood and problem drinking in adulthood. *Addiction*, **93**, 505–514.

Horton, J. H. and Lipsitz, S. R. (2001) Multiple Imputation in Practice: Comparison of Software Packages for Regression Models With Missing Variables. Statistical Computing Software Reviews. *The American Statistician*, **55**, 244–254.

Hupkens, C. L. H., Knibbe, R. A. and Drop M. J. (1993) Alcohol Consumption in the European Community: uniformity and diversity in drinking patterns. *Addiction*, **88**, 1391–1404.

Huxley, R., Neil, A. and Collins, R. (2002) Unravelling the fetal origins hypothesis: is there really an inverse association between birthweight and subsequent blood pressure? *The Lancet*, **360**, 659–65.

Istvan, J. and Matarazzo, J. D. (1984) Tobacco, alcohol, and caffeine use: a review of their interrelationships. *Psychological Bulletin*, **95**, 301–26.

Johnson, F.W., Gruenewald, P. J., Trento, A. J. and Taff, G. A. (1998) Drinking over the life course within gender and ethnic groups: a hyperparametric analysis. *Journal of Studies on Alcohol*, **59**, 668–680.

Kessel, N. and Walton, H. (1989) *Alcoholism.* Penguin Books (second edition).

Kmetic, A., Joseph, L., Berger, C. and Tenenhouse, A. (2002) Multiple imputation to account for missing data in a survey: estimating the prevalence of osteoporosis. *Epidemiology*, **13**, 437–44.

Knibbe, R. A., Drop, M. J., van Ree, M. J. and Saenger, G., (1985) The development of alcohol consumption in the Netherlands: 1958–1981. *British Journal of Addiction*, **80**, 411–9

Korkeila, K., Suominen, S., Ahvenainen, J., Ojanlatva, A., Rautava, P., Helenius, H. and Koskenvuo, M. (2001) Non-response and related factors in a nation-wide health survey. *European Journal of Epidemiology*. **17**, 991–9.

Kuh, D. and Ben-Shlomo, Y. (eds) (1997) *A Life Course Approach to Chronic Disease Epidemiology: Tracing the Origins of Ill-Health from Early to Adult Life.* Oxford University Press.

Kuh, D. and Cooper, C. (1992) Physical Activity at 36 years: patterns and childhood predictors in a longitudinal study. *Journal of Epidemiology and Community Health*, **46**, 114–119.

Kuh, D. and Hardy, R (eds). (2002) *A Life Course Approach to Women's Health.* Oxford University Press

Kuh, D., Hardy, R., Chaturvedi, N. and Wadsworth, M. E. J. (2002) Birth weight, childhood growth and abdominal obesity in adult life. *International Journal of Obesity*, **26**, 40–47.

Kuh, D., Head, J., Hardy, R. and Wadsworth, M. E. J. (1997) The influence of education and family background on women's earnings in midlife: evidence from a British national birth cohort study. *British Journal of Sociology of Education*, **18**, 385–405.

Kuh, D. J. L. and Wadsworth, M. E. J. (1993) Physical Health Status at 36 years in a British National Birth Cohort. *Social Science and Medicine*, **37**, 905–916.

Kuh, D. L., Wadsworth, M. and Hardy, R.(1997) Women's health in midlife: the influence of the menopause, social factors and health in earlier life. *British Journal of Obstetrics and Gynaecology*, **104**, 923–933.

Lahaut, V. M., Jansen, H. A., van de Mheen, D. and Garretsen, H. F. (2002) Non-response bias in a sample survey on alcohol consumption. *Alcohol and Alcoholism*, **37**, 256–60.

Launer, L. J., Feskens, E. J., Kalmijn, S. and Kromhout, D. (1996) Smoking, drinking, and thinking. *American Journal of Epidemiology*, **143**, 219–27.

Lavori, P. W., Dawson, R. and Shera, D. (1995) A Multiple Imputation Strategy for Clinical Trials with Truncation of Patient Data. *Statistics in Medicine*, **14**, 1913–1925.

Ledermann, S. (1956) *Alcool, Alcoolisme, Alcoolisation*, vol 1. Paris: Presses Universitaires de France.

Lee, A. J., Crombie, I. K., Smith, W.C. and Tunstall-Pedoe, H. (1990) Alcohol consumption and unemployment among men: the Scottish Heart Health Study. *British Journal of Addiction*, **85**, 1165–70.

Leigh, B. C. (2000) Using daily reports to measure drinking and drinking patterns. *Journal of Substance Abuse*, **12**, 51–65.

Lemmens, P., Tan, E. S. and Knibbe, R. A. (1988) Bias due to non-response in a Dutch survey on alcohol consumption. *British Journal of Addiction*, **83**, 1069–1077.

Lemmens, P., Tan, E. S. and Knibbe, R. A. (1992) Measuring quantity and frequency of drinking in a general population survey: a comparison of five indices. *Journal of Studies on Alcohol*, **53**, 476–486.

Leon, D. A. (1998) Fetal growth and adult disease. *European Journal of Clinical Nutrition*, **52**, S72–S82.

Leon, D. (1999) Twins and fetal programming of blood pressure: questioning the role of genes and maternal nutrition. *British Medical Journal*, **319**, 1313–14.

Little, R. J. A. (1995) Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, **90**, 1112–1121.

Little, R. J. A. and Rubin, D. B. (1987) *Statistical Analysis with Missing Data*. New York: Wiley.

Longford, N. T., Ely, M., Hardy, R. and Wadsworth, M. E. J. (2000) Handling missing data in diaries of alcohol consumption. *Journal of the Royal Statistical Society*, Ser. A, **163**, 381–402.

Luca, A., Fewtrell, M. S. and Cole, T. J. (1999) Fetal origins of adult disease — the hypothesis revisited. *British Medical Journal*, **319**, 245–249.

Makela, P. (1999) Views into studies of differences in drinking habits and alcohol problems between sociodemographic groups. IVO-Award lecture, Berlin, March 1999. Addiction Research Institute, Rotterdam.

Marmot, M. G. and Brynner, E. J. (1991) Alcohol and cardiovascular disease: the status of the 'U' shaped curve. *British Medical Journal*, **303**, 565–8.

Marmot, M., Ryff, C. D., Bumpass, L. L., Shipley, M. and Marks, N. F. (1997) Social inequalities in health: next questions and converging evidence. *Social Science and Medicine*, **44**, 901–910.

McCleary, L. (2002) Using multiple imputation for analysis of incomplete data in clinical research. *Nursing Research*, **51**, 339–43.

Meng, X.-L. (1994) Multiple-Imputation inferences with uncongenial sources of input (with discussion). *Statistical Science*, **9**, 538–573.

Midanik, L.T (1988) The validity of self-reported alcohol use: A literature review and assessment. *British Journal of Addiction*, **83**, 10191029.

Mulford, H. and Miller, D. (1963) The prevalence and extent of drinking in Iowa, 1961; a replication and evaluation of methods. *Quarterly Journal of Studies on Alcohol*, **24**, 39–53.

Office for National Statistics (1998) *Living in Britain 1996: General Household Survey.* London: HMSO.

Office for National Statistics (1999) Statistics on alcohol: 1976 onwards. *Statistics Bulletin 1999/24*, London: HMSO.

Office for National Statistics (2003) *Living in Britain 2001: General Household Survey.* London: HMSO.

Olsen, M. K., and Schafer, J. L. (1998) A class of models for semicontinuous longitudinal data. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 721–726.

OPCS (1980) *Classification of Occupations.* Office for Population Censuses and Surveys. London: HMSO.

Paneth, N. and Susser, M. (1995) Early origin of coronary heart disease (the "Barker hypothesis"). *British Medical Journal*, **310**, 411–412.

Paul, A. A. and Southgate, D. A. T. (1978) McCance and Widdowson's *The Composition of Foods* (4th ed.). London: HMSO.

Pequignot, G. (1980) Les étapes sur le chemin de la prévention de la pathologie liée a l'alcool. *Toxicomanies*, **12**, 141–154.

Pequignot, G., Tuyns, A. J. and Berta, J. L. (1978) Ascitic cirrhosis in relation to alcohol consumption. *International Journal of Epidemiology*, **7**, 113–120.

Pernanen, K. (1974) Validity of survey data on alcohol use. In *Research Advances in Alcohol and Drug Problems* (eds R. Gibbins, Y. Israel, H. Kalant, R. Popham, W. Schmidt and R. Smart), Vol. 1, 355–374. New York: Wiley.

Pfeffermann, D. (1993) The Role of Sampling Weights when Modeling Survey Data. *International Statistical Review*, **61**, 317–337.

Plant, M. (1997) *Women and Alcohol: contemporary and historical perspectives.* Free Association Books, London/New York.

Poikolainen, K (1985) Underestimation of Recalled Alcohol Intake in Relation to Alcohol Consumption. *British Journal of Addiction*, **80**, 215–216.

Power, C., Rodgers, B. and Hope, S. (1999) Heavy alcohol consumption and marital status: disentangling the relationship in a national study of young adults. *Addiction*, **94**, 1477–1497.

Prescott-Clarke, P. and Primatesta, P. (1998) (eds) *Health Survey for England 1996.* London: HMSO.

Price, G. M., Paul, A. A., Key, F. B., Harter, A. C., Cole, T. J., Day, K. C. and Wadsworth M. E. J. (1995) Measurement of diet in a large national survey: comparison Of computerised and manual coding of records in household measures. *Journal of Human Nutrition and Dietetics*, **8** , 417–428.

Price, G. M., Paul A. A., Cole, T. J. and Wadsworth M. E. J. (1997) Characteristics of low-energy reporters in a longitudinal national dietary survey. *British Journal of Nutrition*, **77**, 833–851.

Reardon, M. L., Burns, A. B., Preist, R., Sachs-Ericsson, N. and Lang, A. R. (2003) Alcohol use and other psychiatric disorders in the formerly homeless and never homeless: prevalence, age of onset, comorbidity, temporal sequencing, and service utilization. *Substance Use and Misuse*. **38**, 601–44.

Rehm, J., Ashley, M. J., Room, R., Single, E., Bondy, S., Ferrence, R. and Giesbrecht, N. (1996) Drinking patterns and their consequences: report from an international meeting. *Addiction*, **91**, 1615–1622.

Richards, M., Hardy, R. and Wadsworth, M. E. J. (1997) The effects of divorce and separation on mental health in a national birth cohort. *Psychological Medicine*, **27**, 1121–1128.

Robinson, R. (2001) The fetal-origins of adult diease: no longer just a hypothesis and may be critically important in South Asia. *British Medical Journal*, **322**, 375–76.

Rose, G. (1992) *The Strategy of Preventative Medicine*. OUP

Rosenbaum, P. R. and Rubin D. B. (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**, 41–55.

Rovine, M. J. and Delaney, M. (1990) Missing data estimation in developmental research. In *Statistical Methods in Longitudinal Research* (ed. A. von Eye), Vol. 1, *Principles and Structuring Change*, New York: Academic Press.

Royal College of General Practitioners (1986) *Alcohol: a Balanced View*. The Royal College of General Practitioners, London.

Royal College of Psychiatrists (1986) *Alcohol: Our Favourite Drug*. Tavistock Publications, London and New York.

Rubin, D. B. (1987) *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley and Sons.

Rubin, D. B. (1996) Multiple Imputation After 18+ Years. *Journal of the American Statistical Association*, **91**, 473–489.

Rubin, D. B. (2000) Software for Multiple Imputation. Discussion paper, http://www.statsol.ie

Rubin, D. B. and Schenker, N. (1991) Multiple imputation in health-care databases: an overview and some applications. *Statistics in Medicine*, **10**, 585–598.

SAS/STAT Software (2003) SAS Institute Inc. htp://www.sas.com

Schafer, J. L. (1997) *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

Schafer, J. L., and Olsen, M. K. (1999) Modeling and imputation of semicontinuous survey variables. *Proceedings of the Federal Committee on Statistical Methodology Research Conference*, 565–574. Washington, DC: Office of Management and the Budget.

Schooling, C. M. (2002) *Heath behaviour in a social and temporal context*. Ph.D thesis. University College, London.

Sheikh, K. (1986) Predicting risk among non-respondents in prospective studies. *European Journal of Epidemiology*. **2**, 39–43.

Single, E. and Leino, V. E.(1997) The levels, patterns and consequences of drinking. In *Drinking Patterns and Their Consequences* (eds M. Grant and J. Litvak), International Center for Alcohol Policies. Taylor & Francis, USA.

Skog, O.-J. (1991) Drinking and the distribution of alcohol consumption. In *Society, Culture, and Drinking Patterns Reexamined* (eds D. J. Pittman, Raskin White),135–6. Rutgers Center of Alcohol Studies: New Brunswick, NJ.

Slymen, D. J., Drew, J. A., Wright, B. L., Elder, J. P. and Williams, S. J. (1994) Item non-response to lifestyle assessment in an elderly cohort. *International Journal of Epidemiology*, **23**, 583–91.

Smyth, M. and Browne, F. (1991) *General Household Survey 1990.* HMSO London.

Spring, J. A. and Buss, D. H. (1977) Three centuries of alcohol in the British diet. *Nature*, **270**, 567-572.

SPSS (1991) *Statistical Algorithms.* Chicago: SPSS Inc.
http://www.spss.com/tech/stat/Algorithms.htm

SPSS (1997) *Missing Value Analysis, Release 7.5.* Chicago: SPSS Inc.

Statistical Solutions (1997) *SOLAS for Missing Data Analysis 1.0.* Statistical Solutions Ltd., Cork, Ireland.

Statistical Solutions (1999) *SOLAS for Missing Data Analysis 2.0.* Statistical Solutions Ltd., Cork, Ireland.

Strachan, D. P., Leon, D. A. and Dogeon, B. (1995) Mortality from cardiovascular disease among interregional migrants in England and Wales. *British Medical Journal*, **310**, 423–7.

Strategy Unit (2003) *Interim Analytical Report for the National Alcohol Harm Reduction Strategy.* Downing Street: Prime Minister's Strategy Unit.
http://www.strategy.gov.uk/files/pdf/SU%20interim_report2.pdf

Sulkunen P. (1989) Drinking in France 1965–79. An analysis of household consumption data. *British Journal of Addiction*, **84**, 61–72.

Susser, M. and Levin, B. (1999). Ordeals for the fetal programming hypothesis. *British Medical Journal*, **318**, 885–886

Taylor, J. M., Cooper, K. L, Wei, J. T., Sarma, A. V., Raghunathan, T. E. and Heeringa, S. G. (2002) Use of multiple imputation to correct for nonresponse bias in a survey of urologic symptoms among African-American men. *American Journal of Epidemiology*, **156**, 774–82.

Thomasson, H. R. (1995) Gender differences in alcohol metabolism: physiological responses to ethanol. In *Recent Developments in Alcoholism* (ed. M. Galanter), **12**, 163–79. New York: Plenum Press.

Tuyns, A. J., Pequignot, G. and Esteve, J. (1983) Greater risk of ascitic cirrhosis in females in relation to alcohol consumption. *International Journal of Epidemiology*, **13**, 53–57.

Uchtenhagen, A. (1990) Alcohol data in longitudinal research. In *Data Quality in Longitudinal Research* (eds D. Magnusson and L. R. Bergman). CUP.

Van Leer, E. M., Seidell, J. C. and Kromhout D. (1994) Differences in the association between alcohol consumption and blood pressure by age, gender, and smoking. *Epidemiology*, **5**, 576–82.

Van Oers, J. A., Bongers, I. M., van de Goor, L. A. and Garretsen, H. F., (1999) Alcohol consumption, alcohol-related problems, problem drinking, and socioeconomic status. *Alcohol and Alcoholism*, **34**, 78–88.

Wadsworth, M. E. J. (1991) *The Imprint of Time : Childhood, History and Adult Life.* Oxford: Clarendon Press.

Wadsworth, M. E. J, Butterworth, S. L., Hardy, R. J., Kuh, D. J., Richards, M., Langenberg, C., Hilder, W. S. and Connor, M. (2003) The life course prospective design; an example of benefits and problems associated with study longevity. *Social Science and Medicine*, **57**, 2193–2205.

Wadsworth, M. E. J., Cripps, H. A., Midwinter, R. A. and Colley, J. R. T. (1985) Blood pressure at age 36 Years and social and familial factors, cigarette smoking and body mass in a national birth cohort. *British Medical Journal*, **291**, 1534–38.

Wadsworth, M. E. J., Mann, S. L., Rodgers, B., Kuh, D. J. L., Hilder, W. S. and Yusuf, E. J. (1992) Loss and representativeness in a 43 year follow up of a national birth cohort. *Journal of Epidemiology and Community Health*, **46**, 300–304.

Whichelow, M. J. (1993) Trends in alcohol consumption. In *The Health and Lifestyle Survey: Seven years on* (eds B. D. Cox, F. A. Huppert and M. J. Whichelow), 237–255. Dartmouth Publishing Co Ltd, Aldershot, UK.

Wiggins, R. D., Ely, M. and Lynch, K. (2000) A comparative evaluation of currently available software remedies to handle missing data in the context of longitudinal design and analysis. Working paper No. 51, Centre for Longitudinal Studies, The Institute of Education, University of London.
http://www.cls.ioe.ac.uk/Cohort/Ncds/Publications/mainpubs.htm

Williams, D., Skinner, R., Silverstone, J., Mann, J., Miller, D., Pyke, D., Garrow, J. S., Hockerday, T., Lewis, B., Pilkington, T., Black, D., James, W. P. T., Besser, G., Brook, C. and Graddock, D. (1983) Obesity: a report of the Royal College of Physicians. *Journal of the Royal College of Physicians*, **17**, 5–65.

Wilson, P. (1980) *Drinking in England and Wales.* HMSO, London.

Wilson, P. (1981) Improving the methodology of drinking surveys. *The Statistician*, **30**, 159–167.

Zhou, X. H., Eckert, G. J., Tierney, W. M. (2001) Multiple imputation in public health research. *Statistics in Medicine*, **20**, 1541–9.

Appendix 1

Weekly Recall and CAGE Questions

SECTION F. DRINKING

1. In the last seven days, how many of the following drinks have you had? (*Do not count non-alcoholic drinks.*)

   Spirits or liqueurs (e.g. whisky, gin, brandy, vodka .....measures ☐☐

   Wine, sherry, martini or port .....glasses ☐☐

   Beer, lager, cider or stout .....half pints ☐☐

2. Have you ever felt you ought to cut down on your drinking? (*Do not include dieting.*)

   Yes ..... 1 ⟶ Have you felt this way in the last year?

   No ..... 0 — Yes..... 1

   No 0

3. Have people ever annoyed you by criticising your drinking?

   Yes ..... 1 ⟶ Has this happened in the last year?

   No ..... 0 — Yes..... 1

   No 0

4. Have you ever felt bad or guilty about your drinking?

   Yes ..... 1 ⟶ Have you felt this in the last year?

   No ..... 0 — Yes..... 1

   No 0

5. Have you ever had a drink first thing in the norning to steady your nerves or to get rid of a hangover?

   Yes ..... 1 ⟶ Has this happened in the last year?

   No ..... 0 — Yes..... 1

   No 0

Appendix 2

The Diet Diary

Instructions (5 pages)

Diary sheets (2 pages) for the first diary day only:

the sheets for the remaining 6 days are copies of these.

D

1989

STRICTLY CONFIDENTIAL

NATIONAL SURVEY OF HEALTH AND DEVELOPMENT

(Medical Research Council)

66-72 Gower Street, London WC1E 6EA
Telephone (01 387 7050)
Extn 5707

DIET DIARY

We would be grateful if you could keep this diary of *everything* you eat or drink over the next five days. The nurse will show you how to keep the diary and leave an example to help you.

As you will see, each day is clearly marked, beginning with the first thing in the morning and ending with food and drink at bedtime. Please treat each day separately. Write in the name of all food and drink taken, a description if necessary and the amount, for each part of the day. If nothing was eaten or drunk during a part of the day, draw a line through that section. Record everything at the time of eating, *NOT* from memory at the end of the day.

Overleaf is a list of popular foods and drinks. Next to each item is the sort of thing we need to know so that we can tell what it is made of and how much you had. This list cannot cover all the foods and drinks that people may have, so try to relate to a similar item if any you have eaten are missing.
For some foods, you may find it easier to describe how much you had by comparing it to one of the pictures.
Many packet foods have weights printed on them, so please use these whenever possible.
At the end are some notes on recording made up dishes and foreign foods.
At the end of each day, there is a list of snacks and drinks that can easily be forgotten. *If not already mentioned in some other part of the day*, please write any extra items in here.

When the last day has been filled in, post the booklet back to us in the envelope provided.

It is *very important* that you do not adjust what you eat and drink just because you are keeping a record. Please stick to your usual diet!

THANK YOU FOR YOUR HELP

Food/Drink	*Description & Preparation*	*Amount*
Bacon	lean or streaky; fried or gilled rashers	number of
Baked beans		tablespoons or tin size or picture 12
Beefburger (hamburger)	fresh or from a packet or take away; fried or grilled; large or small; with or without bread roll	number
Beer	stout, bitter, lager; draught, bottled, low alcohol, homemade	number of pints and half pints
Biscuits	plain; savoury; cheese, crispbread, sweet, chocolate, wafer; home-made; include biscuits like Kit-Kat and Penguin; write in the name if you can	number
Bread (see also sandwiches)	wholemeal, white or brown; currant, fruit/malt; large or small loaf; thick, medium or thin slices; sliced or unsliced	number of slices
Bread rolls (see also crusty or soft sandwiches)	wholemeal, white or brown. Alone or with filling (see sandwiches). Crusty or soft	number of rolls
Breakfast cereal	what sort; cornflakes, weetabix, muesli etc	number of biscuits or tablespoons or picture 1
Bran		tablespoons
Bun	what sort; iced, currants; sweet or plain; large or small	number
Butter for bread	ordinary or low fat	thick, average, thin spread
Cake – small	what sort: cream, iced; sort of filling	number
Cake – large	what sort: cream, fruit, iced; sort of filling	slices, see picture 13 & 14
Cheese	what sort: cream, cottage, hard; low fat; write in the name if you can; large, medium, small helping	tablespoons or picture 2
Chips	large, medium, small helping	see picture 7
Chocolate	what sort; diabetic. Give brand name	number or bar size
Chops	what sort; lean or fatty; large or small; fried, grilled or baked	number
Coffee	with milk; ½ milk/½ water; all milk	cups or mugs
Cooking oil	type; brand name	
Cream	single, double or whipped, low fat; sweetened or unsweetened	tablespoons
Crisps	brand name; low fat; low salt	size of packet
Custard	pouring custard or egg custard	tablespoons
Doughnut	jam, cream, iced, sugared	number
Egg	how was it cooked: boiled, fried, scrambled, poached, omelette, etc	number

Food/Drink	*Description & Preparation*	*Amount*
Fish	what sort: fried, boiled, grilled, poached; with batter or breadcrumbs; in tin with oil or ketchup	helping, see picture 6
Fish cakes or fingers	what sort; large, medium or small size; fried or grilled	number
Fruit – fresh	what sort	number
Fruit – stewed/canned	what sort: sweetened or unsweetened	tablespoons
Fruit – juice	what sort: sweetened or unsweetened	glasses or cups
Gravy	thick or thin	tablespoons
Honey		teaspoons
Ice-cream	dairy or non-dairy; flavour or variety	number or tablespoons
Jam	specify if low sugar	teaspoons
Kidney	pig, lamb, ox; fried or stewed	number or helping, see picture 5
Liver	pig, lamb, ox; fried or stewed	helping, see picture 4
Margarine	soft (in carton) or hard; low fat; give brand name	thick, average or thin spread
Marmalade	specify if low sugar	teaspoons
Marmite/Bovril	what sort	½, ¼, whole teaspoons
Meat pie or pasty	what sort: individual or helping	number: picture 3
Meats	what sort: lean or fatty; how cooked, with or without gravy	slices or helping, pictures 4 & 5
Milk – for drinking on its own or for cereals	full cream, silver top, semi-skimmed, skimmed, sterilised, UHT, flavoured, powered, soya	glasses or cups
Minced beef	on its own: with vegetables	tablespoons or see picture 5
Peanuts	dry roasted or ordinary salted	size of packet
Porridge	with sugar; with milk or cream	tablespoons
Potatoes	baked, boiled, mashed and creamed, fried/chips, instant, roast; with butter	tablespoons: see pictures 10 & 11
Pudding	what sort: eg steamed sponge; with fruit; pie (what sort); jelly; blancmange; mousse: instant desserts, milk puddings	tablespoons or slices or pictures 3, 13 & 15
Rice	brown or white; boiled or fried	tablespoons or picture 8
Salad	describe ingredients, with dressing; what sort of dressing (eg oil and vinegar, salad cream)	tablespoons
Sandwiches and rolls	wholemeal, white or brown bread; what filling: butter or margarine: large or small loaf; thick, medium or thin slices	number of rolls or slices of bread

Food/Drink	*Description & Preparation*	*Amount*
Sauce – hot	(for vegetables, meat or fish; puddings) what sort; savoury or sweet; thick or thin	tablespoons
Sauce – cold	what sort: eg tomato ketchup, brown sauce; salad cream; sweet or savoury	tablespoons
Sausages	what sort: eg pork, beef, pork and beef; low fat; large or small; how cooked	number
Sausage rolls	large or small	number
Scone	what sort: with currants, sweet or plain; cheese	number
Sherry	what sort: eg sweet, medium or dry: at home or in pub	glasses
Snacks – in packet	what sort: eg cheese straws, twiglets, pretzels (give brand name)	packet size
Soft drinks	squash, undiluted or diluted: fizzy drinks; low calorie; give brand name	glasses or cans
Soup	what sort: canned, packet instant or vending machine, homemade	tablespoons, mug
Spaghetti/pasta	canned in sauce, plain boiled	tablespoons or see picture 9
Spirits	what sort: eg whisky, gin, vodka, rum; at home or in pub	single measures as in pub
Sugar	added to cereals, tea, coffee, fruit etc	heaped or level teaspoons
Sweets	what sort: eg toffees, boiled sweets or wrapped (give brand name); diabetic	number
Tea	with/without milk	cups or mugs
Vegetables	what sort: with butter: how cooked or raw	tablespoons
Wine	white, red; sweet, medium, dry	glasses
Yoghurt	what sort: eg with fruit, natural, plain; flavour; low fat	cartons, tablespoons
Made up dishes	what sort: eg vegetable, cheese, fish, meat poultry or mixed, stews; casseroles; dishes made using minced beef such as cottage or shepherd's pie, etc; home made puddings, cakes and biscuits. Please say what the dish is called and give ingredients if you can. Write in the amount eaten in tablespoons, or as a large, average or small portion in comparison to one of the pictures.	
Foreign food	what sort: eg pizzas, Chinese or Indian dishes etc. Please say what the dish is called and give ingredients if you can. Write in the amount eaten in tablespoons or as a large, average or small portion in comparison to one of the pictures.	

Use the pictures to help you to indicate the size of the portion you have eaten. Write on the food record the picture number and size A, B or C nearest to your own helping.

The pictures could also be used for foods not shown, eg, fruit crumble might be similar to shepherd's pie, fruit cake similar to veal and ham pie, and baked beans similar to peas.

1A 1B 1C

Cornflakes

2A 2B 2C

Cheddar Cheese

3A 3B 3C

Pie

4A 4B 4C

Meat

5A 5B 5C

Meat or Vegetable Stew

6A 6B 6C

Fish

7A 7B 7C

Chips

8A

8B
Rice

8C

12A

12B
Baked Beans

12C

9A

9B
Spaghetti

9C

13A

13B
Sponge Cake

13C

10A

10B
Potatoes

10C

14A

14B
Fruit Cake

14C

11A

11B
Mashed Potato

11C

15A

15B
Fruit Crumble

15C

DAY..... DATE.....		
BEFORE BREAKFAST		
Food/Drink	*Description and Preparation*	*Amount*
BREAKFAST		
Food/Drink	*Description and Preparation*	*Amount*
MID MORNING – between breakfast time and lunch time		
Food/Drink	*Description and Preparation*	*Amount*

10

LUNCH		
Food/Drink	*Description and Preparation*	*Amount*
TEA – between lunch time and the evening meal		
Food/Drink	*Description and Preparation*	*Amount*

11

DAY ..... DATE .....

EVENING MEAL

Food/Drink	*Description and Preparation*	*Amount*

LATER EVENING – up to last thing at night

Food/Drink	*Description and Preparation*	*Amount*

BETWEEN MEALS, SNACKS AND DRINKS ***if not already written in before***

Food/Drink	*Description and Preparation*	*Amount*
Chocolate	.....	.....
Toffees/Sweets	.....	.....
Crisps	.....	.....
Peanuts	.....	.....
Other Snacks	.....	.....
Beer	.....	.....
Wine	.....	.....
Sherry	.....	.....
Spirits	.....	.....
Other cold drinks	.....	.....
Tea	.....	.....
Coffee	.....	.....
Other hot drinks	.....	.....
Ice cream	.....	.....
Anything else?	.....	.....
	.....	.....
	.....	.....

Space to write in the Recipe or Ingredients of any made up dishes or foreign food that you have mentioned if not already done above

END OF DAY No. .....

Appendix 3

SOLAS™

Discriminant Method

Manual calculations for simulated dataset 1 (Section 6.6.4.3)

using the SOLAS™ algorithm specified in Section 2.8.4.2.2

SOLAS™ 'Discriminant Method'

The general model on which SOLAS bases its method for "Discriminant Multiple Imputation" for categorical variables is

$$P(Y=y_j \mid X=x) = \frac{P(Y=y_j, X=x)}{P(X=x)} = \frac{P(X=x \mid Y=y_j)P(Y=y_j)}{\sum_v P(X=x \| Y=y_v)P(Y=y_v)} \quad (1)$$

(the last step of which is Bayes's theorem). Writing π_j for the prior probability $P(Y=y_j)$, this becomes

$$P(Y=y_j \mid X=x) = \frac{P(X=x \mid Y=y_j)\pi_j}{\sum_v P(X=x \mid Y=y_v)\pi_v} \quad (2)$$

This expression can be expressed more compactly as

$$P(Y=y_j \mid X=x) = \frac{w_j(x)\pi_j}{\sum_v w_v(x)\pi_v} \quad (3)$$

i.e. as a weighting applied to the prior probabilities π_j, where the weight $w_j(x)$ is defined as

$$w_j(x) = P(X=x \mid Y=y_j) \quad (4)$$

i.e. the conditional distribution of X given that $Y=y_j$.

Data for a set of 1253 completely observed cases randomly selected (according to a non-uniform rule) from 2002 completely observed cases are as follows:

	$X=0$	$X=1$	
$Y=0$	164	479	643
$Y=1$	7	603	610
	171	1082	1253

As estimated from the data,

$$\pi_0 = \frac{643}{1253} = 0.51317$$
$$\pi_1 = \frac{610}{1253} = 0.48683 \quad (5)$$

Since X is a single variate, a one-dimensional Normal distribution is used in this case, with mean and variance for each value of Y estimated in the usual way from the mean and variance of the corresponding values of X.

For $Y = 0$ we have, for the conditional Normal distribution of X,

$$\mu_0 = \frac{479}{643} = 0.74495, \quad \sigma_0^2 = 0.74495\,(1 - 0.74495) = 0.19000 \tag{6}$$

and for $Y = 1$ we have

$$\mu_1 = \frac{603}{610} = 0.98852, \quad \sigma_1^2 = 0.98852\,(1 - 0.98852) = 0.011343 \tag{7}$$

For each value of (μ, σ) the weight $w(x)$ is the value of the density function of the Normal distribution $N(\mu, \sigma^2)$ at x, namely

For $Y = 0$:

for $X = 0$:

$$w_{Y=0}(X=0) = w_0(0) = \frac{1}{\sqrt{2\pi \times 0.19000}} e^{-\frac{1}{2}\frac{(0-0.74495)^2}{0.19000}}$$

$$= 0.21247 \tag{8}$$

for $X = 1$:

$$w_{Y=0}(X=1) = w_0(1) = \frac{1}{\sqrt{2\pi \times 0.19000}} e^{-\frac{1}{2}\frac{(1-0.74495)^2}{0.19000}}$$

$$= 0.77124 \tag{9}$$

For $Y = 1$:

for $X = 0$:

$$w_{Y=1}(X=0) = w_1(0) = \frac{1}{\sqrt{2\pi \times 0.011343}} e^{-\frac{1}{2}\frac{(0-0.98852)^2}{0.011343}}$$

$$= 7.3594 \times 10^{-19} \tag{10}$$

for $X = 1$:

$$w_{Y=1}(X=1) = w_1(1) = \frac{1}{\sqrt{2\pi \times 0.011343}} e^{-\frac{1}{2}\frac{(1-0.98852)^2}{0.011343}}$$

$$= 3.7241 \tag{11}$$

Now we can evaluate the conditional probabilities used in SOLAS for imputation of $Y \mid X = 0$ or $Y \mid X = 1$:

For $X = 0$:

$$P(Y=0 \mid X=0) = \frac{w_0(0)\,\pi_0}{w_0(0)\,\pi_0 + w_1(0)\,\pi_1}$$

$$= \frac{0.21247 \times 0.51317}{0.21247 \times 0.51317 + 7.3594 \times 10^{-19} \times 0.48683}$$

$$= 1 - 3.3 \times 10^{-18} \tag{12}$$

$$P(Y=1 \mid X=0) = \frac{w_1(0)\,\pi_1}{w_0(0)\,\pi_0 + w_1(0)\,\pi_1}$$

$$= \frac{7.3594 \times 10^{-19} \times 0.48683}{0.21247 \times 0.51317 + 7.3594 \times 10^{-19} \times 0.48683}$$

$$= 3.3 \times 10^{-18} \tag{13}$$

For $X = 1$:

$$P(Y=0 \mid X=1) = \frac{w_0(1)\,\pi_0}{w_0(1)\,\pi_0 + w_1(1)\,\pi_1}$$

$$= \frac{0.77124 \times 0.51317}{0.77124 \times 0.51317 + 3.7241 \times 0.48683}$$

$$= 0.17918 \tag{14}$$

$$P(Y=1 \mid X=1) = \frac{w_1(1)\,\pi_1}{w_0(1)\,\pi_0 + w_1(1)\,\pi_1}$$

$$= \frac{3.7241 \times 0.48683}{0.77124 \times 0.51317 + 3.7241 \times 0.48683}$$

$$= 0.82082 \tag{15}$$